AFHRL-TP-81-20

AIR FORCE HUMAN RESOURCES LABORATORY

AFHRL CONFERENCE ON HUMAN APPRAISAL: PROCEEDINGS

Edited by

Cecil J. Mullins

MANPOWER AND PERSONNEL DIVISION
Brooks Air Force Base, Texas 78235

June 1981

Approved for public release: distribution unlimited.

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235

NOTICE

When U.S. Government drawings, specifications, or other data are used for any purpose other than a definitely related Government procurement operation, the Government thereby incurs no responsibility nor any obligation whatsoever, and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

This technical paper was submitted by the Manpower and Personnel Division, under Project 7734, with HQ Air Force Human Resources Laboratory (AFSC), Brooks Air Force Base, Texas 78235. Dr. Cecil J. Mullins (MOAM) was the Principal Investigator for the Laboratory.

This technical paper has been reviewed by the Office of Public Affairs (PA) and is releasable to the National Technical Information Service (NTIS). At NTIS, it will be available to the general public, including foreign nations.

This technical paper has been reviewed and is approved for publication.

NANCY GUINN, Technical Director
Manpower and Personnel Division

TYREE H. NEWTON, Colonel, USAF
Chief, Manpower and Personnel Division

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

14 1. REPORT NUMBER: AFHRL-TP-81-20

2. GOVT ACCESSION NO.: AD-A102 755

3. RECIPIENT'S CATALOG NUMBER

6 4. TITLE (and Subtitle): AFHRL CONFERENCE ON HUMAN APPRAISAL: PROCEEDINGS Held at San Antonio, Texas, 19 — 21 March 1979.

5. TYPE OF REPORT & PERIOD COVERED: Proceedings

6. PERFORMING ORG. REPORT NUMBER

10 7. AUTHOR(s): Cecil J. Mullins

8. CONTRACT OR GRANT NUMBER(s)

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Manpower and Personnel Division
Air Force Human Resources Laboratory
Brooks Air Force Base, Texas 78235

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 16 62102F 17 77341001 19

11. CONTROLLING OFFICE NAME AND ADDRESS
Hq Air Force Human Resources Laboratory (AFSC)
Brooks Air Force Base, Texas 78235

11 12. REPORT DATE: June 1981

13. NUMBER OF PAGES: 211

14. MONITORING AGENCY NAME & ADDRESS(If different from Controlling Office): 12 211

15. SECURITY CLASS. (of this report): Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release: distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

9 Topics include: Application of a Generalized Development Curve to Problems in Human Assessment; Performance Ratings -- Comments on the state of the art; Performance Appraisal -- Some nagging problems and possible solutions; Performance Assessment in Organizations -- Some non-random observations; Public Law 95-454 and Performance Appraisal; and Measurement and Latent Trait Theory.

18. SUPPLEMENTARY NOTES

This technical paper represents the proceedings of a conference on human appraisal sponsored and hosted by Air Force Human Resources Laboratory (AFSC).

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)
aptitudes
criterion
job performance
rating
personnel management

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This technical paper consists of the proceedings from the opening day of a conference conducted 19—21 March 1979 in San Antonio, Texas. The purpose was to bring together several of the researchers who have been concerned with various aspects of problems in human appraisal to exchange ideas and to provide discussion and critique of the directions our respective research efforts are taking. Formal presentations of work and ideas connected with state-of-the-art methods and problems in human appraisal comprise the core of this technical paper. Each formal paper is followed by informal comments and discussion by the participants. The informal materials were taken directly from tape recordings of the proceedings, and, with minor editorial changes by the speakers (who were invited to review their remarks prior to publication) appear just as they were spoken.

(the

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

404415 LB

AFHRL Technical Paper 81-20 June 1981

AFHRL CONFERENCE ON HUMAN APPRAISAL:

PROCEEDINGS

By

Cecil J. Mullins

Reviewed and Submitted for Publication By

LONNIE D. VALENTINE, JR.
Chief, Force Acquisition Branch
Manpower and Personnel Division
Air Force Human Resources Laboratory

This report is primarily a working paper.
It is published solely to document work performed.

FOREWORD

This is a report on a conference of somewhat unusual format. In March 1979, Air Force Human Resources Laboratory (AFHRL) was well along in the conceptual development of an appraisal system for Air Force civilian employees, and field tests of some of the methods and procedures had just begun.

Work on an appraisal system is work on criterion research. Criterion research has remained the most intractable area of psychological research for at least the past 50 years. The typical assault on criterion problems described in the literature begins with energy and enthusiasm, an orderly program outline, and usually some original ideas. After it has been demonstrated that the idea worked (or did not work), suddenly the orderly program is mentioned no more, and the investigator turns his/her attention to other problems. Only a handful of investigators have sustained a vigorous attack on criterion problems for longer than a year or two. Four outstanding examples of such people from the civilian community appear in the following pages, along with a few researchers from the military community. It was decided that it would be of great value to subject the developing appraisal system to the intense scrutiny of some outside experts, to be sure we had not gone down some wrong roads or missed a few right roads in the dim light on the edge of the state-of-the-art. Several other decisions were made at the same time.

1. The conference participants would be a small group of invited members. Not only did budget constraints limit attendance to a small group, but we wanted intensive participation and interplay among all the participants, and this seemed most likely in the cozier confines of a smaller group.

2. We wanted participants with unusual qualities:

a. We wanted high-quality people, as evidenced by their experience in the criterion research area and by their seminal publications in the literature.

b. We wanted people who, even though they had already established themselves as authorities in the field of criterion research, were still open enough to listen attentively to our problems and proposed solutions without "grinding some special axe."

c. We wanted people who would form a congenial group so that discourse and the general flow of ideas might be encouraged by a group attitude of mutual respect among members.

How well we succeeded in attracting people with the above qualities may be judged by the reader. For our part, we feel that we could not have been more content with our choices.

3. The overall progression of the conference was to be from general to specific, in the focus of attention. The first day would be devoted to presentation and discussion of papers in general and theoretical terms, designed largely to sharpen the perception of all attendees concerning the current state of the art. As days passed, the focus would narrow progressively to an exact description of all parts of the appraisal system proposed for implementation, and a formal critique and evaluation by each participant of what we were doing.

4. The scheduling of the conference was also a little unusual. The theoretical papers were requested no later than 5 March 1979, so they could be reproduced and sent back out in sufficient copies and in sufficient time for all attendees to have had ample time to read all papers before the first day of the conference, 19 March. In this way, we could save time by dispensing with the usual reading of papers, and could proceed immediately to discussion. Furthermore, the conference was scheduled in two parts. During the first 3-day meeting, all papers would be discussed and the proposed AFHRL civilian appraisal system would be presented in detail. At the end of the third day, an additional 2-day meeting of the group would be scheduled for about 60 days in the future (actually it turned out to be May 7-8). During this 2-month hiatus, the participants from the civilian community devoted 10 workdays to close scrutiny of the proposed system and to an exhaustive critique of every phase of it. When the group reconvened in May, these critiques were presented and discussed.

This publication is a report of only the first meeting, on 19 March 1979. By its nature, this work is a look, by some of the best people in the field, at the state of the art of human assessment. The final reports (the critiques of the AFHRL civilian appraisal system), which were delivered in May, are planned for publication as a separate document later. The format for this work is fairly obvious. Each chapter begins with a formal paper by one of the participants, followed by additional comments by the author, followed by a critique of the paper by another participant, followed by general discussion by all participants. The informal critiques and discussions of the formal papers are presented more or less as they were recorded on tape with only whatever editing was necessary to clarify meaning. Very little effort was made to put the comments into strict grammatical form, in order to preserve as much as possible the flavor and spontaneity of the remarks.

As the general organizer of the conference and editor of this work, I would like to express my appreciation to each of the participants and my admiration for the work each one did. I would like especially to thank Dr. Wally Borman, whose enthusiasm for such a conference and whose suggestions along the way were simply invaluable. Good work, gentlemen.

LIST OF PARTICIPANTS

Dr. H. John Bernardin
Virginia Polytechnic Institute
Blacksburg VA 24061

Dr. Walter C. Borman
Personnel Decisions Research
Institute
2415 Foshay Tower
Minneapolis MN 55402

Dr. Leland D. Brokaw*
Air Force Human Resources
Laboratory
Brooks AFB TX 78235

Dr. Wayne Cascio
Florida International University
Tamiami Trail
Miami FL 33199

Dr. Michael J. Kavanagh*
School of Management
SUNY
Binghampton NY 13901

Dr. Fred Muckler*
Navy Personnel Research and
Development Center
San Diego CA 92152

Dr. Cecil J. Mullins
Air Force Human Resources
Laboratory
Brooks AFB TX 78235

Col Tyree H. Newton
Air Force Human Resources
Laboratory
Brooks AFB TX 78235

Lt Col Forrest R. Ratliff*
Air Force Human Resources
Laboratory
Brooks AFB TX 78235

Dr. Malcolm J. Ree
Air Force Human Resources
Laboratory
Brooks AFB TX 78235

*These people have changed addresses since the conference:

Dr. Leland D. Brokaw
9715 Carolwood Drive
San Antonio TX 78213

Dr. Michael J. Kavanagh
Dept of Psychology
Old Dominion University
Hampton Blvd
Norfolk VA 23508

Dr. Fred Muckler
Canyon Research Group, Inc.
741 Lakefield Rd
Westlake Village CA 91360

Dr. Forrest R. Ratliff
St. Mary's University
Division of Continuing
Studies
One Camino Santa Maria
San Antonio TX 78284

APPRAISAL CONFERENCE

INTRODUCTORY REMARKS

Dr. Mullins: This conference is sort of an unusual one. We asked all the participants to submit their papers ahead of time, and then when I got all the papers I reproduced them and sent them back out to all the participants so that everyone has read the papers already. Therefore, there's no reason to get up there and read them again. I believe the first thing on our agenda is a word of welcome from Colonel Ty Newton. He is the head of the Personnel Research Division, which has three sections in it. One of the sections is the one in which I'm located, which is doing this kind of work. So, Col Newton, would you please come welcome us all to San Antonio.

Col Newton: This is a pretty good time of the year to be in San Antonio. I'll hold my remarks down and let us get on with the show. We've been interested in ratings for a long period of time, and we've been doing research prior to getting into this concentrated rating effort. Almost 2 years ago, we had a symposium downtown and Dr. Muckler was here. Were any of the rest of you here for that one? We were involved in studying performance ratings and criterion development. Since that time we've gotten into developing a civilian appraisal system for the Air Force. And with that we have really concentrated on ratings. So we want you here to see what we're doing. We want to know your ideas, we want you to give us feedback, and we would like to know what we're doing wrong, what we're doing right, and any changes that we should make. So it's going to be a free-flowing session all the way through. We want comments going both ways and we'd like to have your very best input--criticisms, anything you can give us that will help us out because we're committed to a program now. We don't have any way to turn back. We can't say, "Well, this is too tough, we can't solve it," because we're going to put something out into the field. And what we put out, we would like for it to be the best that we can do. So that is why you're here, to help us in this development, to help us know the right way to go, and to make the right decisions. Come the summer of 1980, we will have our system out in the field, and it's going to affect an awful lot of people; therefore, the rewards will be great and the fall will be hard if we don't do a good job. We're committed and we're into it, and we feel like we're on the right track, but there are a lot of answers that we do not have, and we're looking to you people to give us some of those answers. I will be in and out during this session, and we would like for you to feel at home. If anything is not going right for you, let us know and we'll try to make accommodations. So with that, Cecil, I'll turn it back over to you.

Dr. Mullins: All right, next on the agenda is a keynote address by Dr. Brokaw who is the technical director of the Personnel Research Division.*

Dr. Brokaw: I think that "keynote address" is an awfully formal title for what I'm going to do because that brings to mind visions of political conventions and very talented orators whipping their supporters into a frenzy of enthusiasm to go out and win an election. My oratorical skills are not great and hopefully we shouldn't get too emotional about this problem because we need all the cold rationality we can get.

I'd like to lay an anomaly on you. You know, as well as I do, that on the 19th of March 1979, we are living in a very highly technical, highly industrialized society. Some industries are pouring out a continual flow of things for our use. Airlines are taking us hither and yon, and with the telephone we can talk to anybody we want to, any time we want to, any place we want to. All these things are working, they're all doing a pretty good job, and they all have something in common. Someplace there's some poor little man, some poor little woman, doing a lonesome little job which doesn't amount to much, along with a thousand others, and there's a lead man, or a lead woman, to show how to do what has to be done, and there's a foreman to keep track of the lead people, and there are branch chiefs to take care of the foremen, and there are division chiefs to take care of the branch chiefs, and there are vice-presidents to take care of the division chiefs, and there are presidents to take care of the vice-presidents. In every company there's a hierarchy, people who are in charge, people who are doing the work at various levels, and all these things are working, and they're all working pretty well. Now I'll admit that Laurence Peter is correct. Here and there you find somebody who got one step too high and is really not doing too hot, but I guess the system's big enough to absorb a few of those. But by and large our systems are working and they're working because there are people at various levels of the hierarchy making them work. Now how did they get to those places?

Well, let's turn around and look at personnel psychologists like me and industrial psychologists like some of you, and look at the literature and you'll discover that for 60 years we've been screaming that there's no way to tell how well somebody's doing his or her job. It's obvious people can't do that. You can't evaluate performance, it's too difficult, it's too complicated, it's too far from us. It's obvious that we can't predict performance because look what we've done. We've gone on the job and looked at the job very very closely and we've modified the training that trains exactly for what the job requires. And we've built paper-and-pencil tests and put them in a battery, and we can predict the performance in the schools at a level

*Now called Manpower and Personnel Division of .65 or above. Good predictions. We take our same test, we go on a job, and all we get are .02's and .03's and minus .04's. It's obvious that the measurement on the job is no good. You can't measure what people do on a job! It's an article of faith with us; it's something we take pride and pleasure in. You can't measure what people do on a job!

I submit to you that this can't be. People are measuring what's done on a job because they make this person a lead person and they make that person a foreman, they make that person a vice-president. Somehow they do that and they do it right. Now there must be some way to capture that. There ought to be some way to do that.

Dr. Mullins has scheduled this meeting so we can start from the lofty, the theoretical. It's a good idea, the basic premise of science, the way that we do these things, we sift down through that theory until we get to the appraisal system we're building, gentlemen, to where the rubber meets the road, as the tire manufacturer says. This is the place where we take the theory; this is the place where we take our ideas; this is the place we make them work. Col Newton said that we have a very imposing, very frightening, but at the same time we also have a very magnificent, opportunity.

Now Brokaw, throughout the years, has been known to have a series of laws. And Brokaw's second law is that the principles are simple, but the details will kill you. We're here to look at the details. I think we've got some pretty good principles. Maybe we're wrong. If we are wrong, we want to be told we're wrong. We want you to be complete, and we want you to be candid. We don't want you to be nasty, but we want you to be truthful, and we will behave the same way. We want to know the facts. We want to use these facts to build the best system we can build at this state-of-the-art. Now if that's a keynote speech, you got it.

Dr. Mullins: Thank you Lee, that was an excellent keynote speech. Okay, I think I'm next up.

CHAPTER 1

APPLICATION OF A GENERALIZED DEVELOPMENT CURVE TO PROBLEMS IN HUMAN ASSESSMENT

Cecil J. Mullins
James A. Earles
Forrest R. Ratliff
Personnel Research Division
Air Force Human Resources Laboratory
Brooks Air Force Base, Texas

In a previous paper (Mullins & Ratliff, 1979) we proposed that there is no fundamental difference between predictor and criterion variables. All tests are tests of accomplishment in some area, and all test variables may be used for both predictor and criterion purposes, depending on when they are administered and the decisions to be based on the test scores. If the results are to be used to indicate level of current performance only, and no decisions need to be made concerning future behavior (e.g., to award outstanding performance, to provide counseling for poor performance, or to remove the subject from the area in which he or she is working), then criterion information is sought. If the test scores are needed to aid in deciding whether to move the subject into some other job (e.g., for selection, for promotion to a position involving unfamiliar elements, for transfer, or for special assignment), then predictor information is sought. An important point is that the *same* test score may be useful for *either* purpose.

It should not be necessary to point out that what was said in the previous paragraph about test scores may be said also about ratings, about interviews, and about all other means of assessing people, since tests, ratings, and such are merely different approaches to collecting essentially the same information.

It also seems so obvious as to be almost trivial that accomplishment evolves from simple to complex as people displaying the accomplishment develop. A baby's accomplishments are very simple and relatively easy to measure. Its total repertoire of meaningful responses can be assessed in a matter of minutes, or at most an hour or two. As the baby grows into a child, however, the child becomes rapidly so complex that we rarely attempt to measure its entire universe of meaningful responses and start concentrating on particular relatively stable *patterns* of accomplishment, some of which we call "aptitudes." The *infant's* earlier repertoire begins to separate into broad channels of verbal ability, numerical ability, and so forth, probably due to three broad categories of influence; certain innate physiological propensities, which we may call potential (P), a set of variables which may be called energizers (E), and the necessary opportunities (O) for these variables to interact in the development of the individual.

Still later in the developmental process, it becomes even harder to assess the individual, because the interaction of energizers and opportunity variables brings about still more complicated accomplishment patterns. A young man 25 years old will probably have pursued a few areas rather intensively and will have picked up a modicum of accomplishment in many other areas simply by brushing up against them. He will probably know the rules of at least half a dozen sports; he will be able to make some repairs on his automobile; he will be able to do some carpentry, sewing, plumbing, and electrical work; and he will know at least the basics of our elective system of government. There will probably be several hundred areas in which he has at least rudimentary knowledge and simple skill. At this point, the simpler aptitudinal developments have branched again into many more specific skills; soldering, writing, hammering, sawing, throwing, and on and on. These skills and knowledges branch again later, and combine with each other to produce job competence. For example, soldering skill combined with knowledge of television electronics (and other knowledges) will likely produce a competent TV technician. Soldering skill combined with job-specific knowledge may also produce a jeweler or metal worker. Considering all this it becomes almost inconceivable that any attempt would be made to measure complete overall individual competence of anyone--although the idea is exciting. Even the prospect of measuring competence in one whole job has offered difficulties which have proved insurmountable in the past, largely because the efforts have been restricted, for practical reasons, to testing times of an hour or two, and that simply is not enough time in which to evaluate just those competencies required for the average job. To completely evaluate competence on a single job would probably require 2 to 3 days for the simpler, and perhaps as much as 2 or 3 weeks for the more complicated. Therefore, we ordinarly settle for the one or two tasks performed most often, or those most critical, and hope that competence in our sample of tasks correlates at a satisfactory level with competence on the whole job if we were able to measure it.

The difficulty of measuring whole jobs is also why ratings are used so often for this purpose. There seems to be an implicit belief that a supervisor spends enough time observing the performance of the employee on all facets of the work that the supervisor can provide an accurate assessment of the whole job whereas a test cannot do so. Interestingly, it appears from some of our research that raters seem to do better rating on global, rather than specific, elements of performance.

It is probably true that a very large battery of tests could be constructed and quantified, if management were willing to bear the expense, which could, indeed, provide test assessment on all important aspects of the job, and do so without being influenced by leniency error, halo error, and the myriad other errors which human observers are heir to. But this would be very expensive, and although sooner or later some studies of this type will have to be done, we will probably put it off for as long as possible by continuing to use the much less expensive ratings.

At any rate, I believe that some day we shall have to obtain more complete measurement of the worker as the worker is today, for an overwhelming portion of the variance in tomorrow's competence is determined by today's proficiency. Figures 1, 2, and 3 show graphically the kind of development we have been discussing.

At this point, it becomes important to specify what is meant by "developmental level." This term refers to whatever skills, knowledges, and behavior patterns the individual has achieved and now carries around, like a bag of tools, inside the nervous system. It is a very similar idea to what we ordinarily call memory content, if we permit the term "memory content" to carry skills such as sports, soldering, instrumental music, and hammering in exactly the same way it carries knowledges of algebra, chains of command, science, and literature. The curves in the figures have been oversimplified, and make a basic assumption which is obviously untrue--that the relative influences of P, O, and E remain constant throughout the individual's lifetime. Each curve is computed from the formula, $D_{t2} = D_{t1} (1 + i)^{t_2 - t_1}$, where D = level of development, t = some point in time, and i = an increment determined by some interaction of potential (P), energizers (E), and opportunity (O). What the formula says is that an individual's development during any time period in his or her life is a function of previous level of development, the interaction of P, O, and E, and the length of time during which P, O, and E can operate.

Potential refers to very basic innate individual differences which probably remain relatively stable over time, similar to Horn's concept of "analage functions" (Horn, 1968), Cattell's "fluid intelligence" (Cattell, 1941), and Hebb's "Intelligence A" (Hebb, 1941). Potential refers to qualities much more basic than aptitudes, which appear later and are probably rooted in whatever these basic potential variables may be. In this sense, aptitudes may be thought of as accomplishments, much the same as, say, a proficiency test in mathematics.

Opportunity refers to the amount of exposure to conditions favorable to development of a skill or knowledge. It includes formal exposure to training, education, and experience, and less formal, rather accidental exposure to people and situations from which one is likely to learn. It also includes hobbies, discussions, and casual reading.

The word "Energizers" refers to all those variables that impel the subject to act or to pay attention, or that keep him or her doing so longer than other people. This would include differences in native physiological energy, those variables which energize differentially called interest and motivation, and those more general ones such as value structure and dedication.

P, O, and E interact in many obvious ways. If you have a natural knack for some behavior (P), it is likely that you will be rewarded in some way for displaying it, increasing the likelihood that you will seek or create (E) other conditions (O) in which the display can be repeated. The practice makes your knack still more impressive, which increases even more the likelihood that you will find the activity rewarding, which in turn . . .

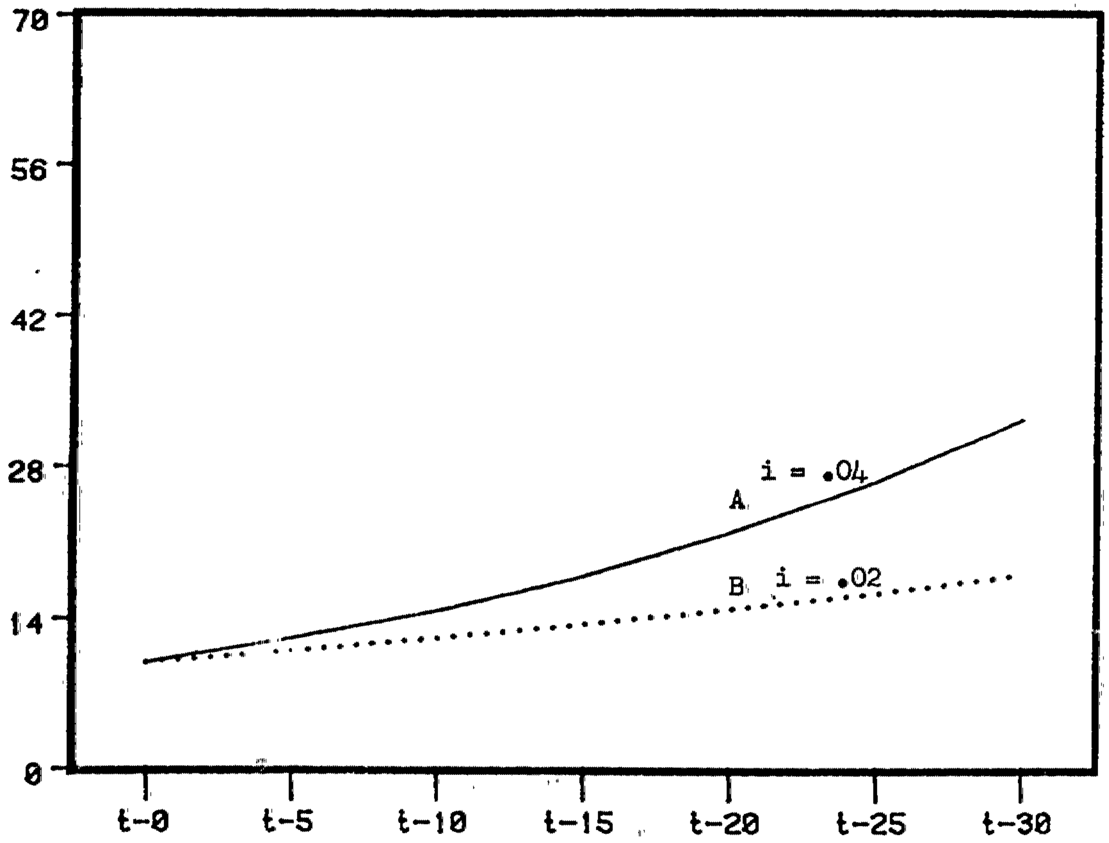
It is time now to return to the curves in Figures 1, 2, and 3. It should be mentioned very early that the shape of these curves was chosen to represent development as a whole person. It would probably also apply to very broad job areas. If one is thinking about the development of a particular skill or a specific knowledge, the curves would likely involve cubic functions and be S-shaped, since psychological limits would be approached relatively quickly. That does not seem so likely with whole jobs, however. A welder probably learns the specific art of welding as a brief S-shaped function, but constellations of learning surrounding the work of welding and defining the job of welder go on through an entire lifetime, such as the devising of special jobs and hold-downs, pricing of time and materials, knowing where to obtain supplies and equipment, time-saving shortcuts, and so on.

Furthermore, it would be folly to pretend that the exact nature of the interplay of P, O, and E, to form i, is known. We don't even know yet exactly what the qualities subsumed under P, O, and E are, although we have made some guesses. Nevertheless, we believe strongly that the development curve of most people will follow the general form of the previously stated equation, since this line conforms to what we know about the growth of people.

The intervals on the horizontal time line (t_0, t_5, t_{10}, . . .) are arbitrary numbers, and the range of these intervals is equally arbitrary. For example, t_0 can equally well denote the moment of birth or the 25th birthday. The lines merely represent a range of interest. The intervals on the vertical axis are also arbitrary, and assume that total development of the individual can somehow be measured (i.e., across his or her entire collection of knowledges).

Figure 1 depicts graphically the rapidly accelerating separation of development rate, as time passes, between two individuals whose levels at the first observation are equivalent, but whose i-terms are different. Remember that i is determined by an interaction of P, O, and E. Therefore, a deficiency in any one of these terms should result in the lower line, unless the deficiency happened to be compensated for by an uncommonly fortunate endowment of one of the other two terms. Conversely, uncommonly high P, or O, or E should produce the higher curve, if the lower curve is taken to be normal. The important point to be noticed is that a very small difference early in developmental level becomes a very great difference later, if there is a constant difference in any of the three components of i. If t_0 is assumed to be time of birth, it seems reasonable to assume that, since O and E cannot have played much of a role before birth, P must be approximately equal for these two individuals, and the divergence of the two lines must represent essentially what happens if one is born into a family of a sharecropper and the other into the family of a college professor.

$D_{t_2} = D_{t_1}(1 + i)^{t_2 - t_1}$ $i = \dot{P} \leftrightarrow O \leftrightarrow E$



	t-0	t-5	t-10	t-15	t-20	t-25	t-30
A	10.0	12.2	14.8	18.0	21.9	26.7	32.4
B	10.0	11.0	12.2	13.5	14.9	16.4	18.1

TIME LINE

Figure 1. Development Curve, Differences in i only

$$D_{t_2} = D_{t_1}(1 + i)^{t_2 - t_1} \qquad i = P \leftrightarrow O \leftrightarrow E$$

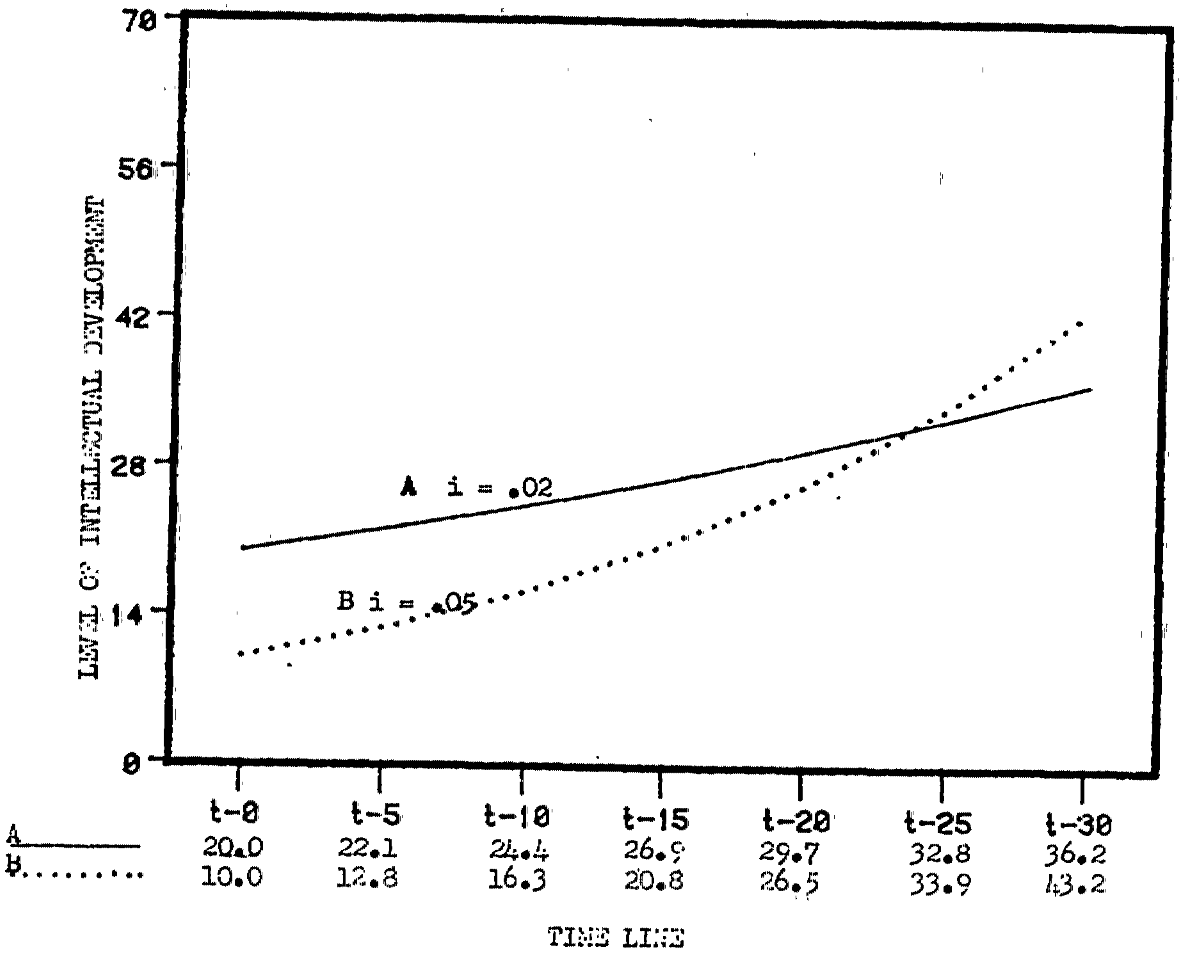


Figure 2. Development Curve, Differences in i and t_0

$D_{t_2} = D_{t_1} (1 + i)^{t_2 - t_1}$ $i = P \longleftrightarrow O \longleftrightarrow E$

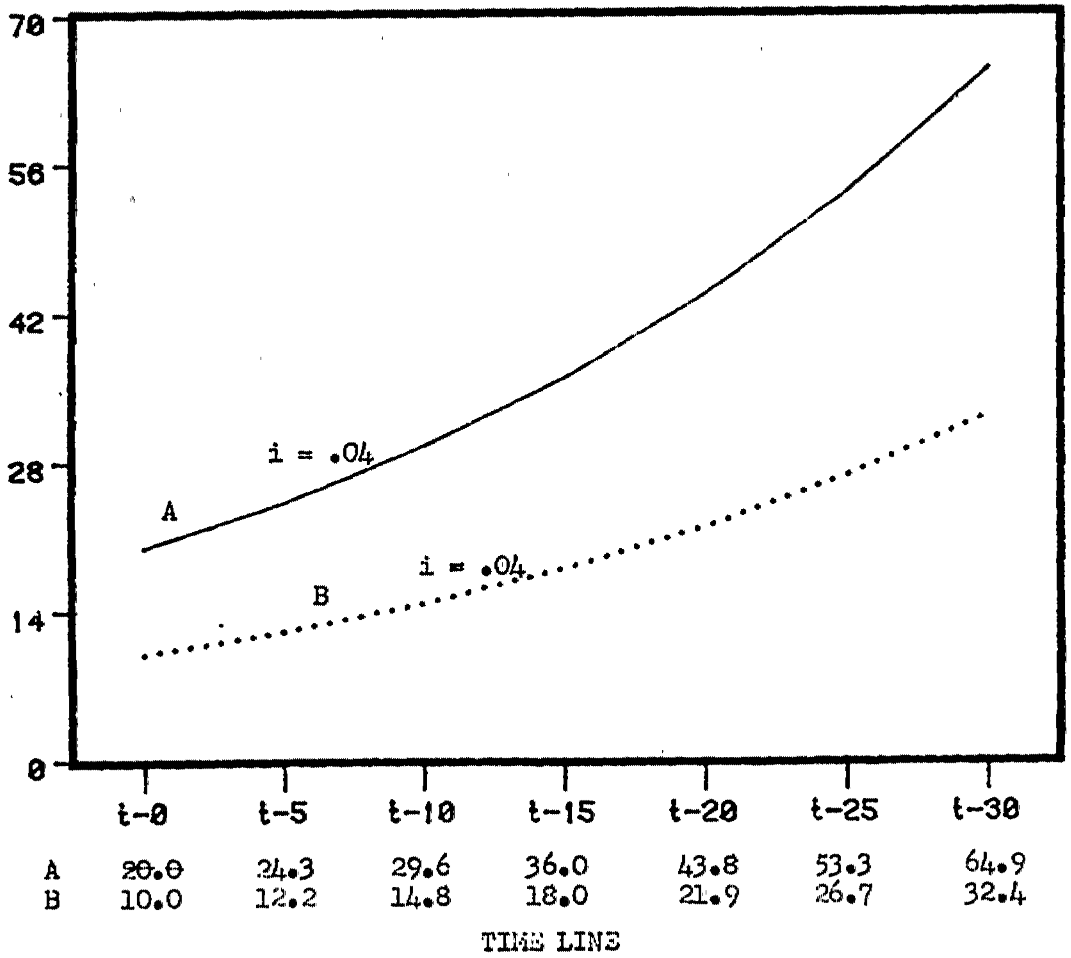


Figure 3. Development curve, Differences in t_0 only

Using the same folksy example, and again assuming that t_0 = birth, Figure 2 shows what happens if the sharecropper's child has twice as much P as that of the professor's child, but much less O or E. If the individual with the upper curve had just a little more O or E (say, i = .03), the other individual would not have caught up in this range of interest.

On the other hand, if t_0 represents some time after birth, and if something has artificially raised person A to the beginning point we see in the curve without permanently changing i, Figure 2 shows how vigorously person B would pass person A later on.

Figure 3 illustrates an interesting comparison of two lines, both with i = .04, but beginning at different levels of development. A difference of 10 points at t_0 becomes a difference of 32.5 points at t_{30}. The lower line in this figure would never catch up to the upper line--indeed, the difference becomes larger and larger with time. If t_0 = birth, Figure 3 states that, given O and E, the individual represented by the top curve will steadily pull away from the other. Since we have assumed that t_0 = birth, Person A must have considerably more P than Person B, but enough less O and/or E that both i-terms are the same (.04). In brief, the simple difference of elevation at the beginning observation point means a considerable difference in developmental level later.

We wondered what i the lower curve would have to have to remain parallel with the upper curve. As it turns out, there is no constant i which will produce points the same distance from the upper curve (say, 10 points) along the entire range of interest. One can compute the i for any t_x point with the formula

$$(1)\quad i_2 = \sqrt[t]{2(1.04)^{t}-1} \quad - 1 \qquad \text{or, more generally,}$$

$$(2)\quad i_2 = \sqrt[t]{\left(\frac{D_{1t1}}{D_{2t1}}\right)\left(1 + i_1\right)^{t} - \left(\frac{D_{1t1} - D_{2t1}}{D_{2t1}}\right)} \quad - 1.$$

where t is the difference between t_1 and t_2

i_2 is the i for the second curve, to be found

D_{1t1} is the developmental level of the upper curve at t_1

D_{2t1} is the developmental level of the lower curve at t_1

and i_1 is the i for the upper curve.

Using formula (1) and Figure 3, an i of .0746 is required to produce a developmental level on the lower curve at t_5 exactly 10 points below the developmental level at t_5 on the upper curve. Analogously, the i needed to produce a developmental level on the lower curve 10 points below the one on the upper curve at t_{10} is .0696. The i required to produce a developmental level at any t_x point on the lower curve 10 points below the developmental level on the upper curve decreases with time, but never becomes quite so small as the i of the upper curve (.04, in this instance).

This formulation again emphasizes the importance of the interaction among i, the passage of time, and the early achievement of a high level of development. It takes considerably more change in i to effect sudden growth than to reach a comparable level of development after the passage of more time.

This curve indicates another interesting point which bears directly on measurement research. If we allow t_0 to be, say, 40 years of age for Individual A and 20 years for Individual B (A would have had an accomplishment level of 10 about 20 years earlier, the same as B), not only would B never catch up with A, but B would not even maintain the same distance behind A. This seems not only to illustrate the reasonableness of considering age as an important predictor of accomplishment, but also argues for the desirability of the earliest possible training in desired behaviors and for the usefulness over a lifetime of a temporary increase in E (hard work), enough to move one's accomplishment up a significant notch.

Focusing on the characteristics of any one of the lines, some more interesting speculations occur. It is rather obvious that the best predictor of any point on the line is the nearest possible previous point on the line. The message of this observation is that the best predictor of performance is the most recent estimate available of past performance. That should come as a surprise to nobody, but a corollary of this platitude seems worth mentioning. It has already been said, or at least implied, earlier in this paper that testing is done most efficiently on simpler concepts, and that ratings appear to be more efficient on global concepts. Therefore, unless one is prepared for a very expensive test evaluation, it is probably better to seek good ratings of recent total job performance in a prediction situation than to use any other measurement method available under the current state of the art. It would probably be better to collect the best possible ratings, for example, of recent job performance than to use the usual aptitude tests, since i becomes a more and more important consideration as time passes between the point when aptitudes develop (in childhood) and the time when the performance occurs which one is trying to predict. This by no means is intended to derogate the utility of aptitude testing. It is certainly much better than no assessment at all, and it is better than ratings carelessly or naively collected. It does mean, however, that much more complete testing by much more complex means than usual should be employed, or one should go to good global ratings, even with their multitude of known defects. This conclusion, in turn, points to the desirability of more and better research on ratings, for poor ratings are not useful for any purpose. At the present time, we are just beginning to devise means for studying accuracy of rating scores, not to mention the difficult problems of learning what to do with sets of rating scores varying in their accuracy from ratee to ratee, from rater to rater, and from situation to situation.

These curves, if they prove to be useful at all, should help us in organizing our assaults on the human development problem, which includes the problem of more complete assessment of humans. Human measurement, in turn, includes a great number of other problems such as appropriate methods of collecting data on the most appropriate variables at the most appropriate points in the worker's life.

The curves keep telling us that "intelligence"--in practice if not in concept--is not immutable and is not the only important vehicle which moves accomplishment. Indeed, in this view, the only intelligence which can be measured at any important time after birth is a trait which has been developed in the person by an interaction of P, O, and E. The concept of intelligence as it can be measured later in life is most nearly exemplified by what we have called "developmental level." All we can measure is functional, as opposed to native, intelligence, and potential is only one component of it.

In a practical sense, the distinction is not of great importance. The curves also tell us that if one individual enjoys an efficiency advantage over another, <u>for whatever reason</u>, the second individual will have to work harder, or receive more opportunity, or be blessed with more innate potential, or all three, in order to catch up. If all other things are equal, the individual who has developed to a higher level at a given time will find still further development proceeds faster, simply because of the elevation.

Most people who have thought about these problems have probably made all these observations for themselves. The only advantage provided by the curves is that they help systematize and explain the phenomena and hopefully may lead to new insights into the development and display of human abilities.

BIBLIOGRAPHY

Cattell, R.B. Some theoretical issues in adult intelligence testing. Psychological Bulletin, 1941, 38, 592 (Abstract).

Cattell, R.B. Theory of fluid and crystallized intelligence: A critical experiment. Journal of Educational Psychology, 1963, 54, 1-22.

Hebb, D.O. Clinical evidence concerning the nature of normal adult test performance. Psychological Bulletin, 1941, 38, 593 (Abstract).

Hebb, D.O. The Organization of Behavior. New York: Wiley, 1949, 294.

Horn, J.L. Organization of abilities and the development of intelligence. Psychological Review, 1968, 75(3), 242-59.

Mullins, C.J., & Ratliff, F.R. Criterion Problems. Chapter XI in Criterion development for job performance evaluation: Proceedings from symposium. AFHRL-TR-78-85. Brooks AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory, February 1979, 98-109.

Dr. Mullins: You've all read my paper. I've copied the charts on these transparencies. I don't think I'll need to show all three of them. I'll just put one of them up there and leave it while we talk and if one of the others strikes your fancy, I've got all three of them here to show.

I don't have too many remarks to add to what was said in the paper. The basic idea, as I'm sure you saw, is a kind of theoretical concept which may be useful for getting some indications as to where we need to focus. In a sense this is a model looking for an application. A few of the applications that we've already thought about are fairly general and fairly broad. Mostly this concept, I think, explains a few things that have been puzzling me over the years.

One of these is the idea of what's meant by an intelligence quotient, or what's meant by an intelligence score. If you look upon this curve of development as development of whatever, including intelligence, then I think we can see why so many confusing studies have appeared in the literature. Take the old heredity/environment controversy, for example. There's some excellent evidence indicating that intelligence is almost entirely heredity, and there's some excellent evidence indicating that it's almost all environment. I believe the chart makes it clear that intelligence is merely an expansion of competence of a particular time, and that it must be affected by things that happen in one's life, and that, if that is true, these things must happen to something innate because life does begin at some point. Given this, it seems to follow that intelligence--and other competencies--should be expected to be stable but not immutable. It has nothing to do with something which is set, something that can never change, but at any particular time, whatever a person displays as intelligence is what has been achieved given the things that person started with. Added to that idea is the basic notion, somewhat like Hobart's apperceptive mass, that the higher you are on this line at any given moment, the easier it is to go higher, so that things come in accumulating powers.

Of course, that has very little to do with what we're trying to accomplish here but I'm trying to give you some general applications of this idea. It would also explain some of the confusing results they've had with the Head Start program. I thought this was one of the best things ever conceived until I saw the initial data from it. The initial data indicated that it was doing nothing. I couldn't believe that, but later some other evidence came in that had been taken a little earlier after Head Start. The first data came later on, say 2 or 3 years later. And it seemed that these children were right back where they started. Later data came in which were taken a little earlier in the game, say 6 months after Head Start was over, and it looked as if it had done a great deal to bring these children up. I think that the explanation can be found in this general approach to achievement and to intelligence, to performance, however you want to name the concept. If you don't substantially change any of the interactive bits (potential, opportunity, and energizers), if none of those are changed permanently, then we expect, I think, for that line to go up sharply whenever the opportunity is there, but then drop back again when the opportunity is removed. Does this make sense?

I think, narrowing down a little more to the practical, to me this changes somewhat my idea of what the criterion problem is. The criterion problem is no longer to me a separate problem from a predictor problem. It just simply changes where you look at this line in terms of a person's development. If you're looking from a particular point in time backward, then you're looking at a criterion situation. If you're looking from a particular point in the line forward, you're looking at a prediction situation. And so far as I can tell, that's about all there is that is different between the two.

The best predictor of future performance, and this is so well known it's practically trivial, the best known predictor of future performance is past performance. But I think, judging from this and from some of the evidence that seems to support this, that the best predictor of all is past performance taken as recently as we can get it. And all that seems to be embodied in this concept.

Now that brings on some further complications and leads us perhaps to something else which is important in the practical situation, and that is that to get the most recent past performance to use as a predictor, you also have to get the most complicated kind of measurement that you can possibly get.

It's very difficult to measure everything that a person knows, everything that a person has accomplished at any particular time of life unless it's at birth. Consequently, if you're trying to predict say from the age 35 to the age 36 it becomes imperative that you get some extremely complete and complex measures at the year 35.

In general, our research and most of the research in the field has indicated that this is very difficult to do with tests, but not as difficult to do with ratings. Ratings seem to be about as good whether you're measuring a global concept or a specific concept and so this might indicate to us where you need to apply ratings as opposed to applying testing. Except for those two applications, I think, as I say, this is a model looking for a use. The best use we've made of it is to give us a few clues into what kind of research we ought to be doing.

I think it should also be equally obvious that we really don't know what we're talking about, at least on the detail level. For example, we don't have any idea what that interaction function is between P and O and E. We don't even know that this is the actual form of the curve that we have up there in the formula but we think it's something along this general line. This curve fits so well what we know about what we've seen in the field that it seems to me it has to be some curve similar to that. But the terms in the exponent may change, the terms in the one-plus-i may change, perhaps some of them turn negative. It seems likely that may be true because there are instances where people, instead of gradually going up, sometimes go up and back down and then back up again. But we think that whatever curve eventuates from this will be at least along this general form of line. Whether or not that formula is correct and whether or not we know the exact interactions of P, O, and E, still we find it useful and we find that this does fit the data that have been collected.

That's about all that I had to add to the paper and I'd like to turn it over now to Dr. Borman to discuss my paper.

Dr. Borman: Let me make a few comments that I've prepared beforehand, and then we could move into a full discussion maybe of the comments and of your paper, if that's a reasonable format.

I had a couple of observations on the paper. The first one was, I would appreciate learning more about how you formulated this particular equation. Frankly, my first reaction was that this was a very arbitrary kind of equation, so I set about trying to develop a simpler equation that might be as reasonable conceptually. And the result of that was, I couldn't. My conclusion to that was that you put in quite a bit of thought and effort in developing that particular equation. I may be totally wrong, this may have been an afternoon drinking session or something where you came up with this, but I would be very curious to know more and learn more about how you developed this particular equation because I'm sure that we'd all benefit from that kind of development effort.

Another thing that struck me in the paper is your comment about global ratings being, in a sense, better than ratings of specific elements of performance. This really struck me because I thought I was the only one who was bothered by this kind of thing. In some of the data that we've collected, one in a large-scale Army study where we collected both ratings of Army units on very specific behaviorally oriented scales and also on a final scale--it was overall morale, that had no behavioral anchors, no guidance whatsoever (those were the two kinds of rating stimuli the raters used). We found that the interrater agreement, the interrater reliability was much higher for the latter kind of rating scale where they had no guidance whatsoever but were to make this very overall, global kind of assessment. And also in three different Navy recruiter studies that I've conducted where we've had not only behaviorally anchored scales but also trait kind of dimensions that were very carefully defined, and also dimensions based on multidimensional scaling analysis that we had quite a bit of confidence in, and an overall performance kind of dimension; we really had much higher interrater agreement again for the latter kind of scales. So I've seen the same phenomenon in data that we've collected and have done some thinking about it, and I think it would be worthwhile discussing that.

It seems to me that first of all maybe we're simply gathering data that reflect a kind of commonly held bias--that individuals when faced with these behaviorally anchored or other more rigorous kinds of dimensions pay attention to the stimuli very closely, and for some reason have some trouble actually making their ratings based on those stimuli. But when we turn them loose and ask them to make overall kinds of judgments, then everybody knows what we're talking about. And it may be, again, that we are capturing some non-valid kind of variance in the form of a commonly held bias when we collect global ratings. On the other hand, it may be that we're actually asking a more reasonable question of raters. It may be that people think more normally, more naturally, in rather global terms, and some other data that we've collected suggest that some dimensions are in fact much easier to use than others. And I'm talking now about the content of dimensions where certain kinds of interpersonal dimensions may be easier for individuals to rate very accurately, but you give them another kind of dimension and consistently they cannot rate those kinds of dimensions as accurately. So I think this is really worth looking into because it has a lot to do with the kinds of dimensions that we present to raters. You know we may be really barking up tne wrong tree with this behavior scaling idea, for instance.

Let me just make a pitch for ratings which actually Cecil was making. In my experience in trying to gather objective kinds of measures of performance on a number of different kinds of jobs, I've been constantly frustrated by the criterion deficiency problem. It seems to me that almost no matter what kind of job you're talking about or dealing with, in trying to gather objective kinds of performance data, even if you're a good snooper and have some very good ideas about how to evaluate performance using some kind of objective measures, it seems to me, conceptually the measure always comes up short--and I mean far short. I think this is one place where raters can really help us out a lot. As long as we can define the performance domain in a reasonable way by developing different dimensions of performance, I think a rater has, at least potentially, an extreme advantage over any kind of objective measure or even series of objective measures. This is just a different way to look at ratings; it's a more positive kind of way of looking at ratings.

I would not argue that ratings as they generally are practiced presently have this kind of advantage, but I think potentially they have a really magnificent advantage, if we can tap this kind of advantage. I'd kind of like to throw this open to Cecil because I don't know whether this is reasonable at all. Maybe this is the way that my mind works but I was trying to think of some ways that this model could be tested. I realize it's quite premature to actually plug numbers in or evaluate individuals on the P, O, and E variables and track them over time to see what kind of development actually occurs in individuals, but it would be a noble effort, I would think, to try to do that on a limited score; try to essentially fit the curves for a few individuals and see whether there was any kind of consistency there and then perhaps move to change the formula around if it were not actually fitting a number of individuals. I wonder if you have done any thinking about that, Cecil, or what do you think? That it's absolutely ridiculous at this point to actually collect data to try to test this kind of thing?

Dr. Mullins: To take the last one first, certainly we've also thought of this and we've even thought of some ways to attack it. But there is simply no way, I think, in our current situation that it could be done. In the first place, it's not important to the problems that we have. The only place that I can see in this whole model where you can apply any leverage is at birth, because at birth the O and the E would be minimized if not completely eliminated. There would be some E, I suspect, having to do with nourishment of the mother prenatally and that sort of thing; the O at birth, I suspect, would be zero. So at birth the major difference among people would probably be in the P. And if you could ever latch on to one of these interactive elements (P or O or E), and get it pretty well defined and observe its effect later, then perhaps you'd have some lever for attacking the others. I don't believe that this model is anywhere near ready for testing yet, to answer your question. I have no idea of the relative importance of O and E and I suspect that probably they interact even more than they do with P. I think P certainly is affected by opportunity and energy and vice versa, but I think that the interaction between O and E is far greater than it is between P and either of those two. But these are just simply gut hunches. I don't know, and I think a great deal of work would have to be done to find out. I think also that P is going to turn out to be something, as I mentioned in the paper, far more primitive than what we ordinarily call aptitude. It's going to be something much below that--perhaps the speed of an impulse along a neuron, or perhaps a tendency to clump things in perception. But I believe that if one did have the opportunity to experiment with very young people, I'd say 2 years and under, this might be the place to begin.

Dr. Cascio: Is this why you said age is an important predictor of accomplishment? That's what you said in the paper.

Dr. Mullins: No. The reason I said that is that just the fact that you've lived for a certain period of time, unless you're a carrot, means that you have learned something. It may not be directly applicable to your job, but I'm not sure that there's anything which isn't applicable to your job, at least indirectly. Given a choice between two people equally trained on a given job, one of whom also has a lot of other information--automobile mechanics, electronics, whatever; it makes no difference--and the other one who has nothing except what he's been trained to do the job, I'd much prefer the former. Things will come up where knowledge in other areas will be, if not directly transferable, at least so in principle.

Dr. Cascio: I'd like to jump on another idea that Wally brought up about the global ratings apparently capturing more than simply objective criteria. Last year in the Journal of Applied Psychology, I published an article called "Relations Among Criteria of Police Performance." We looked at almost 1,000 road patrol officers, 36 criteria, 8 behaviorally anchored rating dimensions and 28 measures of on-the-street behavior, and after an awful lot of massaging of those data, we found that objective measures seem to capture less than 20% of the variance in overall job behavior. I think that what that says to me is that those things are deficient in some way. This was a very comprehensive study and we looked at what we figured was every possible indicator of on-the-street behavior, and yet when you look at those global ratings, they seem to capture a lot more than what's actually objectively obvious.

Dr. Bernardin: I think, in terms of the global ratings, that the whole problem is biased as a systematic source of variance and when you use it as a dependent variable which we call interrater reliability, you're going to inflate it with halo, or with the way I discussed in my paper, illusory correlations. It's natural that that will be inflated, and I think that's the reason you're going to get higher interrater reliability with global ratings. Thinking more naturally in terms of global kinds of traits, that disturbs me, because that's the way the literature seems to be going. There's a recent article by DeCotiis (OBHP, 1978) that compared very specific behaviorally anchored rating scales (BARS) to global ratings with police officers, and another article by Schneier in Proceedings of the Academy of Management (1979) getting back to a global or trait approach. When we talk about thinking more naturally, in terms of global ratings, I think you're talking about physical attractiveness. If you want to talk about natural variables, physical attractiveness, height, weight, sex, race, those are natural variables. I don't think we can deal with those kinds of variables. Bias is going to inflate those kinds of dependent variables. Also, and this of course isn't even getting into where the rubber meets the road in terms of Equal Employment Opportunity Commission (EEOC) and stuff like that, I just don't think that the research or the direction of performance appraisals should go in that direction because I think it's simply either illusory correlation or halo.

Dr. Kavanagh: One of the things that Wally said, and my reaction to the issue of global and specific, is that we may have made some false assumptions as industrial psychologists. We have a truism that we will have better ratings, the more specific they are and the more clearly we can break down the behavior. In our lives we find the reality is that we are usually making decisions on the basis of overall or global kinds of information. If we decide to marry, if we decide to go on a date, if we decide to promote, if we decide to hire, they're largely binary decisions. I think that people in our organizations think that way, think administratively, rather than growth and development. So I believe it's easier to think globally, and people think that way, and maybe we've been trying to enforce our behavior-specific ideas on the way we want people to think.

Dr. Borman: I really agree that we can't move in the direction of going as far as height, weight, physical attractiveness, and so on, but clearly a test of this is to compare not only with the interrater agreement but also the validity of these kinds of judgments. The only data that I know of, thinking about it quickly, is assessment center data from the Campbell, Dunnette, Lawler, and Weick study which, as I recall, actually show that overall estimates of potential--which is a very global kind of rating--I believe, were found to be more valid than the specific, individual, more behavioral kinds of dimensions.

Dr. Bernardin: The argument I would make to that is the Klimoski and Strickland argument (Personnel Psychology, 1977) about a common bias in the criterion that was used on the job.

Dr. Kavanagh: I think there's another major point and this is in relationship to the model. I think that Wally brings up a good point--how do you test it? Well, P, O, and E are obviously vectors and there are models to test them if you can specify the variables. But I think the other point that's made is the concept that has been lingering in the literature, and that is the idea that a criterion is a dynamic phenomenon. Thus, the nomological net should include this dynamic aspect, and that somehow we, in our practical world, should be able to get that into our measurement systems. We don't do so at the present time. And going out on a limb, I suggest that considering global versus specific, maybe the reason for the apparent superiority of global is that we aren't capturing the dynamic nature. If you could capture the dynamic nature, I have a feeling that the prediction and the understanding from the more specific kinds of behavioral dimensions would be much better. Lee Wolins began working on this problem 15 years ago, and sponsored a dissertation that argued that essentially we should be predicting future behavior from quadratic functions over time. In a couple sets of rating data, I factor analyzed the quadratic functions and I got something, but I can't figure out what it is yet. As most of us know, prediction from childhood intelligence is terrible, but Wolins did sponsor one study of childhood intelligence, in which they were able to fit the typical linear model and added a significant quadratic function to it. That's what I like about this concept. I think that's the value of this developmental idea, that there is change going on, and if there is change, it might be a lawful change. If it's a lawful change, we should be able to measure it. I don't know yet how we'd include it in our performance appraisal systems. There are systems that try to include it; e.g., Management by Objectives (MBO) systems.

Dr. Mullins: I believe when we get into the system that we're going to describe to you that we made for going out into the field, you'll find that we've incorporated much of MBO and perhaps a few ideas beyond it.

Dr. Kavanagh: I have a 1975 article here by Bill Scott and Clay Hamner from Organizational Behavior and Human Performance (OBHP) where they look at variations of performance profiles, and I'll read you some of the findings. They experimentally manipulated two variables. They had high and low variability in terms of worker performance over 47 time trials. In some cases they had people starting at a high level and descending in terms of their performance over the time trials, and others were increasing. The findings that I feel are important are: (1) that they demonstrated that the high variability workers were judged to be more able but less motivated; and (2) they found the workers showing improvement were rated as more motivated than those whose pattern of performance was either random or deteriorating over time. This makes sense. If you're going to select someone into a graduate program, and two applicants have a B average, where one has showed an increasing trend over the 4 years of college and the other a decreasing trend, the former student is probably going to get the nod if you only have one space. I wonder how much impact these order effects, these time effects, and change in performance have on performance rating, and how we can incorporate them within a system. Your particular model indicated to me that this is an important point.

Dr. Mullins: Now to respond specifically to what you asked me. You asked where the formula came from and said that it showed considerable thought in developing. Actually where it came from, we were sitting around the office one day and for some reason we were talking something economic. And then the idea of education came up and somebody mentioned that education was a lot like an investment. So I had had my set on economics and then this came up and I got to thinking that probably there is some curve of development that fits pretty closely the return of interest on an investment. That's essentially what this is. It's just a basic little old return of interest on an investment formula and I simply took variables that looked reasonable and plugged them in, and that's it. So that's how much thought went into it. A lot of thought has gone into it since then.

Now for the global versus specific conversation. I think we sort of strayed off the idea that I wanted to get across here. The basic consideration here, it seems to me, is that people develop from simple to complex and this complex keeps getting more complex and more complex and so on. Whenever you get to the point where someone is already a full-fledged adult, I believe that the only way you can really measure his or her effectiveness at that time, completely measure it, would be to design certain tests that we don't even have yet and then use perhaps 3 weeks to 3 months of testing time to apply them.

I personally am prejudiced in favor of tests and against ratings because we've used ratings so much in the past and they've been so disappointing. Perhaps they'll be better in the future.

However, we still have the practical problem of administering that many tests and designing that many tests and it is so much easier to get a rating. Now that leads us into the next point that you made and that was the rating of global versus specific aspects. I think I've already mentioned to you that we've done about three or four studies which keep pointing very solidly in the direction that people seem more comfortable rating a global concept. This doesn't necessarily mean traits. It means how does he or she do on the job as opposed to making widgets, and how does he or she do changing tires, and how does the worker do any specific sort of thing. Just an undefined global rating of how good this worker is, according to the studies we've done, appears to do as well as you could if you analyze specific behaviors and try to predict something else with the global plus these other things. The specific ratings just don't seem to add any predictive variance.

So all that we've done leads us to suspect, and I believe in my paper that's all I said was we suspect--"evidence indicating that perhaps," this kind of hedging. But the work that we've done indicates to me that probably there is very little you can get from a rater, from the unsophisticated rater, beyond how good this guy is at what he's doing. Now if that's so and if the rating is good enough to do some prediction as the global ones seem to be in the work we've done, then it seems just practical to take a global rating from a rater for that very complicated phase in the model where otherwise you'd have to test for 3 or 4 or 6 weeks.

There was one more thing that I wanted to say relative to that, concerning Wally's comment that I was making a pitch for ratings. I am and I'm not. There are trade-off disadvantages and advantages for considering tests as opposed to ratings. I much prefer tests because if they're good tests they're objective. There's no argument about how the guy performed. But ratings have advantages, too. If you have something exceedingly complex in nature, probably your best bet is to go with ratings. If you have something fairly simple like aptitudes or something of that type where you intend to predict over a long number of years, probably you'd do better with tests. That's all the pitch I was making.

Dr. Bernardin: One more thing about that "exceedingly complex." Isn't that contradicting what you said earlier about the specificity of the information? If you can only rate on global traits or we can only do good things with global kinds of ratings, doesn't that interfere with the study of complex processes?

Dr. Mullins: I don't follow you.

Dr. Bernardin: You mentioned the ratings are good for complex kinds of studies, looking at complex relationships, but you said earlier that global ratings seem to be doing the job and ratings of specific kinds of variables are not succeeding. Wouldn't you need the latter in order to study complex processes?

Dr. Mullins: I don't see why. If you're using the rating as a criterion measure or something of that type, we can find no advantage using several specific ratings over using one global rating. I think that a global rating is an extremely complex measure in itself, and consequently probably matches the structure of the job somewhat better than these uni-factorial ratings that one usually collects. Have I missed your point?

Dr. Bernardin: No, that's it.

Dr. Muckler: Could we see that figure once more? I'm not uncomfortable with that function, and favor that function for the younger work force, but I'm thinking in terms of the above 55-year-old work force, and now with the loss of the mandatory retirement I think we can still address those problems here. Your equation, I think, can handle any problem. But what we're seeing in the older work force is that they function somewhat differently.

Dr. Mullins: Perhaps a leveling off?

Dr. Muckler: Well, in some parts but in others certain kinds of decrements are occurring. And some of us who have to deal with that problem are rather eagerly watching the aging literature because it looks like decrement is reasonably specific but you don't get a general fall-off. First you start losing certain kinds of things, but it's not sure what you're losing. The literature just isn't that substantial. But the way the labor force is working, plus this loss of mandatory retirement, I think we're going to see an awful lot of people from 55 to 75 attempting to perform in the work force, and I think something is going to look considerably different.

Dr. Kavanagh: If there are decrements, that would imply a maximum point. The question that I have is where does the maximum point come from, and where does the decrement come from? As Cecil said, are we looking at potential as a given that really doesn't change greatly? Is it opportunity or is it energizer mostly that might cause change?

Dr. Muckler: The ones I've seen I think are energizers. What I see mostly in the older work force is, I hate to use the word, but it's really a decrement in motivation and frankly, although my sample may be a strange one, an increasing bitterness about the system. I'm getting an awful lot of really deep hostility from the older members of the work force. I see it I think more among the production people. The older people are very much frightened of what their old age is going to be like. In terms of content, I hear more discussion about that. They're really, for example, concerned with the social security system. Most of what I hear in content are survival questions. Am I going to survive? I was really quite serious about this. I know an older guy who is 67 and now he doesn't have to retire and he's going to stay as long as he can walk in the door and he is looking forward to senility because he feels he can really screw things up.

Dr. Kavanagh: The reason for my question was because I was thinking of it from exactly the other end. I was thinking about the young worker who had entered the work force within the last 5 or 10 years. Those people perceive that the opportunity is either not there or they are overtrained for the opportunity that's there. So I argue that that artificial maximum point occurs even though potential is greater and energizers are there, but opportunity is simply not there. My point is that, if we can identify this sort of situation, then we can start talking about management interventions. The logical outcome of what we just said is different performance assessment systems for different cohorts or age groups.

Dr. Mullins: There was one point that I wanted to make about the shape of the curve. Certainly if you're talking about some specific narrow area of performance, like repairing automobiles, or doing mathematics, or whatever, I think probably the curve would be S-shaped to approach maximum performance and then it would level off. Possibly if it was a very boring job, it would drop back again, or something of that type. But remember that this curve is supposed to picture the entire, total development growth of the individual. Probably as people reach the maximum in one area and find that they can't proceed there any further they'll go and take up some new sport or hobby. Anything that the individual has achieved, all of that goes into this curve. And I'm not sure that in one lifetime one would ever level off.

Dr. Muckler: In the aging literature on IQ it seems to me each new paper that comes out extends the time period further and further. You see each new paper shows less and less decrement. I'm not sure which one to believe, currently, but they're now showing pretty stable IQ up through age 70.

Dr. Bernardin: There was a great exchange in the American Psychologist a few years ago where one person maintained there are major decrements at a fairly early age and he developed this, and the rebuttal was "so and so was obviously practicing what he preaches."

Dr. Muckler: One of the things that I see over the spread of people that I supervise and observe is some really fundamental different value systems now operating, between the folks who are in their 20's and the folks who are in their 50's and 60's. I'd really like to explore a little bit what the value system we're dealing with is. The 20's and 30's folks now seem to have much less identification with the system, with the organization, much less commitment to it, and much more willingness to move than I've seen before.

Dr. Mullins: I think that's a fascinating observation. I think it's a basic insight into some of the problems that we may ordinarily have with our prediction system and as you'll see tomorrow or the next day from Mr. Wilbourn, we have a big prediction study coming up soon which will involve around 16,000 subjects, plus another 44,000 cohorts, and it might be that we might take a look at trying to develop some separate prediction equations by age, as well as by level.

Dr. Borman: Excuse me, I want to ask a question about that approach. Are you saying that there should essentially be different management systems, different ways of handling the different cohorts, or are you saying in addition the performance rating scale in some way, or the things that people are evaluated on, should be different? Because if it's the latter, I'm not quite sure about that. It seems to me if we have a job, you know there are certain job requirements. I'm sure it is true that for a lot of complex jobs there may be different ways to succeed on the job and perhaps it would be good to allow somehow on the rating form for these different ways to succeed. Primarily in those different age groups perhaps fairly uniformly succeeding or failing in different ways. And that will be very interesting if you can actually develop different forms for different cohorts.

Dr. Cascio: I'd like to bring up an extreme example because that question really occurred to me last summer. Don't all laugh at once. I was developing BARS for garbage collectors. Solid waste employees. Well, I was working for a county government and they had all different kinds of jobs and we were looking at the performance appraisal systems for lots of different classes of employees from bus drivers to these solid waste collectors, to managers, and so forth. And I remember talking with people in the solid waste department and they were saying, "Okay now, what are you going to do with a 67-year-old garbage collector who is just marking time? Are you going to appraise this person on the same things as everybody else and how are you going to use those data? Are you going to use it for personal development, are you going to use it for promotion? What are you going to do with it?" I'm not sure; I don't have the answer. I raised the question because it set me to thinking last summer that maybe we need different types of appraisal systems for different age cohorts, different groups of people. I don't have the answer for it. It's an extreme example that illustrates the problem.

Dr. Mullins: I find one thing fascinating about what you just said. What do you promote a garbage collector to?

Dr. Cascio: Driver. He's not a toter anymore, he's a driver.

Dr. Mullins: He's the executive.

Dr. Cascio: And then you have the scheduler. If you're an expert driver then you become a scheduler and you send people out on different routes.

Dr. Mullins: It never occurred to me that a garbage collector was ever going to go anywhere except to the dump.

Dr. Cascio: Well, if the collector doesn't do well as a toter, then demotion can occur also--to somebody who just sweeps up in the dump after everybody else has dropped off the trash. It's an extreme example but there are promotional ladders and demotional ladders as well.

Dr. Bernardin: And my argument would be why have a performance appraisal system at all for that type of person, a lame duckish type?

Lt Col Ratliff: I think that everybody is entitled to being appraised, for part of motivation and satisfaction is a day-by-day thing in terms of your relationship with supervisors and work crew members.

Dr. Mullins: I think there's probably a need to be evaluated now and then even if you're only marking time. That's just a guess.

Dr. Kavanagh: I'm going to go back a bit and respond. I don't see management's use of performance assessment information and the performance assessment system itself as independent. Unfortunately they are oftentimes in our organizations independent and treated independently. The management system should determine the type of performance assessment, and the performance assessment system clearly impacts both upon the management evaluation and use of that information. I think that we are looking at different measurement reasons for different age cohorts. For the younger worker, we are talking about successful performance on the job. With the younger worker we're expecting mobility. We're expecting movement; we are also trying to assess whether or not we should keep that individual within our organization. If we make a decision to keep that individual within our organization, then we must make a decision on where to move that individual; that is, career progression. The standard answer is to move everybody we keep into supervision, which may be why supervision is in such terrible shape. There are other answers, and many corporations are seeking these in terms of career planning and placement kinds of programs tied to the assessment of individuals. This would enable me to say when I assess you: "You're doing a fantastic job; you really can't do any better on this job. But you'll never move into <u>my</u> job because you don't have this set of experiences and you must move laterally to this other job." In an organization in which I've been developing a performance assessment system for about the last 2 years (painfully injecting it), there are people who will never be formally appraised again, to reinforce John's point. And these same people are told on a day-to-day basis that they are doing okay. They get good feedback, but there's just no formal appraisal. There's no formal appraisal because it just doesn't serve any purpose at this point; people are not going to be transferred to different jobs. For example, one lady is going to retire as personnel director in 3 years. They're getting good feedback, and that's an appraisal. That's what Cummings and Schwab called the old MAP approach--you know, the maintenance of current performance approach--you've got to keep their motivation going. So there are different reasons for performance appraisal for employees of differing ages. I don't know of any system that has captured that. It's a complex problem.

Dr. Mullins: It's now a little after 9:30 and it's time for Dr. Borman's presentation. Okay, Wally.

CHAPTER 2

PERFORMANCE RATINGS: COMMENTS ON THE "STATE OF THE ART"

Walter C. Borman
Personnel Decisions Research Institute

Paper Prepared for AFHRL Conference
on Human Assessment

19 March 1979

This paper has three parts. First, it reviews current knowledge related to improving performance rating accuracy or validity; that is, it discusses what we have learned about the effects of rater, rating format, rater training, and administrative factors on the accuracy of performance ratings. The second part examines possible limitations in the precision with which persons can make performance ratings, and argues that a significant breakthrough in reducing rating errors awaits more knowledge about the rating process. The third part then outlines one approach toward beginning to study the performance rating process.

1. Approaches to Increase Rating Accuracy

This section of the paper examines five different strategies for increasing the accuracy of performance ratings: (a) selecting raters likely to provide accurate ratings; (b) ensurirg that raters have good opportunity to observe ratee behavior relevant to the performance dimensions; (c) employing a rating format that aids raters in making accurate judgments of ratee performance; (d) training raters to make more error-free ratings; and (e) ensuring that the rating "set" or situation encourages raters to provide relatively accurate performance ratings.

Throughout this section, I will focus on rating accuracy[1] or validity as a dependent variable rather than on psychometric criteria such as leniency, halo, or restriction of range. This is because accuracy appears to be the most critical measure of error in ratings. Accuracy provides "the final word" on the quality of ratings while other psychometric errors are in a sense substitute or indirect measures of accuracy. Of course, one would think that the extent of leniency, halo, or other such errors in ratings should correspond closely to the level of accuracy in those ratings, but certain results suggest this is not necessarily the case (Borman, 1975, 1977; Borman & Rosse, 1978; Buckner, 1959; Crow, 1957; Freeberg, 1969). This further suggests that much of the research employing these indirect measures of accuracy may have provided misleading results and that, wherever possible, we should focus on rating accuracy as a dependent measure.

[1] Unless otherwise noted, "accuracy" refers to Cronbach's differential accuracy (DA; Cronbach, 1955). DA for a rater is the correlation between his/her ratings of three or more ratees on a single dimension and some kind of criterion "true scores" assigned to the ratees on the dimension. DA has been termed the "purest" measure of accuracy (e.g., Hastorf, Schneider, & Polefka, 1970; Sechrest & Jackson, 1961) because it avoids various response biases.

Selecting raters: Individual differences correlates of rating accuracy. A preliminary question that must be asked here is how consistently are individuals accurate (or inaccurate) in rating performance? That is, if a person provides relatively accurate ratings in one setting, how likely is it that this person will provide accurate ratings in another setting? A reasonable degree of consistency is obviously required if we are to study individual differences correlates of rating accuracy.

Research and speculation by psychologists studying person perception suggest that consistency in interpersonal accuracy across situations is quite low. For example, Crow and Hammond (1957) found near zero correlations in accuracy across rating tasks, and Gage and Cronbach (1955) [and, more recently, Dunnette (1968)] have argued that we should not expect high across-rating task consistency when rating situations vary so greatly. The type of person being rated, the relationship between rater and ratee (e.g., friend, acquaintance, stranger), and the nature of the rating judgment (along trait dimensions, on a behavior checklist, etc.) are just some of the ways the rating situation can vary, and the existence of a generalized ability to perceive others accurately across the wide range of situations possible does appear unlikely.

However, within a relatively narrow range of rating contexts, we may expect more consistency. Mullins and Force (1962) found that airmen were reasonably consistent in accuracy when estimating their peers' "carefulness" and vocabulary test scores. And Borman (1977) found moderate consistency (intraclass r = .46) in performance rating accuracy across two different jobs. Perhaps within the kinds of situations raters typically face in appraising performance there exist reasonably stable individual differences correlates of rating accuracy.

What might these individual differences correlates be? A recent study we conducted (Borman, 1979) addressed this question. In the study, 16 scripts describing persons performing on two jobs--recruiter interviewer and manager--were prepared in such a way that the performers' effectiveness on various dimensions of performance approached a preset, realistic level. Five- to 9-minute performances of these scripts were videotaped, and "true scores" of effectiveness were developed by obtaining expert ratings of performance on each relevant job dimension. One hundred forty-six college students then completed a series of inventories tapping various individual differences and rated the performers' effectiveness on each dimension. Differential accuracy (DA; Cronbach, 1955) scores were computed for each subject, and DA scores were correlated with inventory responses. The most consistently high correlates of accuracy were intelligence, personal adjustment, and detail orientation.

In general, results of this study are remarkably similar to certain results from person perception studies; that is, studies that have as the rating task perception of others' personality, the prediction of others' opinions, etc. Table 1 depicts this correspondence between studies, suggesting that at least some individual differences are important for successful interpersonal perception across a variety of settings. These findings tend to support a "general-specific" hypothesis offered by Borman, Hough, and Dunnette (1976), one which posits that certain individual differences are consistently related to interpersonal accuracy across a variety of situations, while other individual differences correlate with accuracy in only specific kinds of contexts.

It appears then, that some raters are simply consistently better than others at providing accurate portrayals of ratee performance. The study discussed above (Borman, 1979) found that approximately 17% of the variance in rating accuracy is accounted for by individual differences. This is a significant "chunk" of variance, suggesting that rater individual differences do indeed contribute significantly to performance rating accuracy.

Raters' position to rate. Only a few studies have examined the effects on accuracy of raters' position in relation to the ratee. Many studies evaluate the effects of rater position on interrater reliability, leniency, or halo, but again, such studies may be misleading because of the poor correspondence sometimes found between accuracy and other psychometric properties. In the most directly relevant studies, Freeberg (1969) found that raters with opportunity to view ratee behavior relevant to three cognitive abilities provided more accurate evaluations of these abilities than did raters with little relevant contact with ratees. And, Whitla and Tirrell (1953) found that supervisors closest in organizational level to a group of ratees provided ratings that correlated higher with job knowledge test scores than did the ratings of supervisors more distant in terms of organizational level.

In addition, studies that consider the predictive validity of ratings suggest that relevance of the behavior observed to the performance being rated is important for obtaining accurate ratings. For example, Amir, Kovarsky, and Sharan (1970); Hollander (1965); Tupes (1957, 1959); and Waters and Waters (1970), among others, have found that peer ratings successfully predict subsequent performance. The main reason advanced for this result is that peers are typically in good position to observe work behavior relevant to future performance requirements, in better position than supervisors or others with relatively little opportunity to observe ratee work behavior.

One refinement of this position regarding peer ratings is that peers may not have the best view of the entire performance domain; supervisors, subordinates, or other groups may be better qualified than peers to evaluate some facets of performance (Borman, 1974; Campbell, Dunnette, Lawler, & Weick, 1970). Accordingly, Borman suggested that raters be subgrouped into homogeneous clusters in terms

Table 1

Correlates of Interpersonal Accuracy in Previous Studies
and in the Borman (1979) Study

Previous Studies	Borman (1979) Study[a]
. Intelligence (Taft, 1955; Vernon, 1933; Wedeck, 1947)	. Verbal reasoning (.31)
	. Intelligence, high grades, and investigative interests (.26)
. Dramatic and artistic interests (Estes, 1938; Taft, 1955)	
. Adjustment; emotional stability (Green, 1948; Scodell and Mussen, 1953; Taft, 1955)	. Freedom from self-doubt and disillusionment; acceptance of others; tendency not to worry or become stressed (.26)
. Good impression; sense of well being; tolerance; self-control neuroticism (-) (Vingoe and Antonoff, 1968)	. Self-control (.20) . Tolerance (.19)
. Sense of well being; tolerance (Hjelle, 1969)	. Empathy (.17) . Aggression (-.17)
. Social detachment; independence (Adams, 1927)	. Social interests; interest in observing others (-.17)
. Achievement via conformance (-) (Edwards and McWilliams, 1974)	. Detail orientation (.18) . Self-perceived detail orientation (.24)
. Task versus social orientation (Taft, 1955)	. Heterosexuality (-.18) . Exhibition (-.17)

[a] Variables are included in this part of the table only when they correlate with overall DA at the .05 level or greater.

of their relationships with ratees. These separate rater groups might then provide ratings using only dimensions appropriate to their expertise and position to rate. An empirical finding in Borman's (1974) study which supported this approach was that raters provided more reliable ratings on dimensions for which they seemed to be in good position to make judgments of ratee performance. Some evidence supporting this refinement (but focused on a validity argument rather than on a reliability argument) is provided by Buckner (1959) and Einhorn (1972).

First, Buckner (1959) had members of three different organizational levels (enlisted men, noncomissioned officers, and officers) rate the job performance of their enlisted submariner shipmates. He found that the more disagreement there was in ratings across croanizational level, the more successful a composite of their judgments was in predicting relatively complex criteria. It may be that when the three levels disagreed, they were accurately rating relatively uncorrelated aspects of the ratees' performance.

Second, Einhorn (1972) had four judges predict survival time of persons who had contracted Hodgkin's disease. Raters were asked to judge each patient's state of health on nine components related to the disease and to make an overall prediction of each patient's survival time. Einhorn found that a cross-validated multiple correlation using as predictors of survival time the component and global ratings of all four judges (40 predictors) was higher than any of the individual judge's multiple correlations. The higher cross-validated R for this condition appeared to be partially a function of different judges providing valid information on different components.

Just how important this refinement is remains unclear. But certainly both research results and common sense argue strongly for the importance of raters being selected according to their opportunity to observe work behavior relevant to the performance being evaluated.

Format effects on rating accuracy. Many studies have been conducted to determine the effects of rating format on psychometric properties of ratings. Format comparison studies have examined the relative psychometric "goodness" of forced choice and graphic rating scales and, more recently, of behaviorally anchored rating scales (BARS), summated scales, scales anchored with adjectives or simply numbers, and Guttman-type behavior scales. General conclusions from this research are: (a) limited psychometric superiority is evident for BARS (e.g., Bernardin, 1977); (b) the added time and expense required to develop the relatively sophisticated BARS format is probably not worth it if psychometric considerations are of primary interest (Dunnette & Borman, 1979; Landy & Farr, 1980; and Schwab, Heneman, & DeCotiis, 1975); and (c) rigor in scale development rather than choice of a particular format may be the important consideration in developing rating formats (Schwab, Heneman, & DeCotiis, 1975; Bernardin, 1977; Bernardin, Alvares, & Cranny, 1976).

Again, studies focusing on psychometric properties such as leniency, halo, and restriction of range are very important, but more critical are the effects of rating format on *accuracy* in performance ratings. A study we just completed (Borman & Rosse, 1978) did focus on accuracy.

Briefly, a training program found previously to reduce a number of rating errors was adapted for use in the study (more later about the training part of this study), and five rating formats that have shown promise in helping to reduce rating errors were developed for two jobs. As in our previous research, the ratees were actors performing on two jobs in 5- to 9-minute videotape segments. One hundred twenty-three college student subjects were assigned randomly to 1 of 10 cells defined by the five format and two training treatments (trained vs. untrained), and raters used the format assigned them to evaluate the effectiveness of the videotaped ratees' performances. Significant format effects were found for accuracy, but a strong job x format effect suggested that no one format is consistently "best" across situations.

Table 2 presents the mean accuracy scores for the four different formats and indicates first that the magnitude of the effects is small (w^2 = .03 and .02 in the ANOVAs for the two different jobs). And second, the more sophisticated behavior oriented formats clearly do not yield more accurate ratings than the simple numerical form. We may conclude therefore that BARS not only fail to enhance substantially the psychometric qualities of ratings, but also they do not appear to increase raters' accuracy in rating performance. In fact, *in general*, rating format has proven to have little effect on rating behavior. Perhaps we should heed the advice provided by Landy and Farr (1980) to place "a moratorium on format-related research."

Training raters to make more accurate ratings. At least some of the zeal in studying rating formats has recently been diverted to the study of training effects on psychometric error. It makes good sense that training raters should reduce rating errors, and a number of researchers have now designed training programs and evaluated their impact on psychometric errors. To summarize, rater training has typically been effective in this regard. For example, Brown (1968) successfully trained raters to reduce halo; Taylor and Hastman (1956) found that a treatment in which individual attention was given to supervisor raters during the rating task resulted in lower scale intercorrelations (halo); and Borman (1975) used a short lecture on halo to reduce that error, though lower interrater reliability resulted. In an attempt to reduce *leniency error* in performance ratings, Levine and Butler (1952) designed a group discussion training program and found that participants did indeed provide ratings with reduced leniency. Latham, Wexley, and Pursell (1975) studied the effects of training on a number of rater errors. Their workshop treatment which provided participants an opportunity to practice observing and rating actual videotaped ratees sharply reduced contrast, halo, similar-to-me, and first impression rating errors.

Table 2

Mean Accuracy Scores for Trained and Untrained Raters on Each Rating Format

	Manager Job		Recruiter Job		
	Trained	Untrained	Trained	Untrained	Overall Average
Numerical	.77[a]	.77	.83	.78	.79
Behavior Summary [b]	.79	.78	.74	.75	.77
Summated	.68	.71	.78	.77	.74
BARS	.77	.75	.73	.69	.74

[a] These accuracy indices were derived by computing mean z transformations of the Pearsonian correlation coefficients used to compute DA and then transforming the mean z's back to correlations.

[b] This is a behavior oriented form whose scales are anchored with statements summarizing the content of successfully retranslated behavioral examples.

A somewhat different approach to training raters was used by Bernardin and Walter (1977). They asked college student subjects to keep diaries of their instructors' teaching performance and to use this information to make detailed performance ratings at the end of the term. Ratings by persons who had kept such diaries contained less leniency and halo error than the ratings made by students who had not kept diaries.

Finally, a study by Bernardin (1978) examined effects of a short training session patterned after Borman's (1975) and an hour-long program that provided more in-depth training in avoiding rating errors. Halo and leniency error were reduced, especially for persons who experienced the 1-hour program; however effects on the training dissipated after 13 weeks.

A recent study by Borman and Rosse (1978) focused directly on the effects of training on rating accuracy. A variant of the program developed by Latham, Wexley, and Pursell (1975) was employed to train raters to overcome certain rating errors and ratings made by the trained raters were compared to those made by untrained raters. Results showed that although training reduced halo somewhat, accuracy was not enhanced. This is an intriguing (albeit discouraging) finding. Let's speculate for a moment on an implication of this result. Consider training in relation to three different *classes* of rating criteria. The first class involves what we might term "rating behavior." Leniency, central tendency, halo, and restriction of range are examples of this type of criterion. The second class of criteria relates to interrater agreement and includes convergent and discriminant validity and interrater reliability. The third class involves accuracy and encompasses the various accuracy components discussed by Cronbach (1955), Cline (1964), and others, the most conceptually appropriate for performance ratings being differential accuracy (DA).

Studies exploring training effects on ratings have typically measured success of training on criteria of the first class, and changing rating behavior might be considerably easier to accomplish than improving interrater agreement or accuracy of ratings. Directing persons to "spread out their ratings" or to "provide fewer high ratings," for example, is relatively straightforward, but teaching them to rate more *reliably* or more *accurately* may well be more difficult.

As an aid to speculating about the kinds of training that may be more successful in this regard, let's review a simple three-step model[2] of the rating process (Borman, 1978). These three steps are: (a) observing work-related behavior; (b) evaluating each of these behaviors; and (c) weighting these evaluations to arrive at a single rating on a performance dimension.

Now what kinds of training might increase *interrater agreement* in performance ratings according to this view of the rating process? First, training focused on standardizing the observation of behavior would be important. Second, the model emphasizes the importance of teaching raters a common nomenclature for defining the organizational or societal relevance of the behavior which is observed. (For example, a frame of reference for defining the performance effectiveness levels of different job behaviors should somehow be provided to raters.) Third, interrater agreement should be reached regarding the relative importance of different kinds of behaviors as contributors to effective performance. .

Further, to increase *accuracy*, it is apparent that these agreed-upon effectiveness levels for individual behaviors and the weights assigned to behaviors in developing a final picture of performance effectiveness for a dimension should be "correct," uncontaminated by factors irrelevant to performance-related considerations. It is

[2] This "model" is admittedly much more restricted in scope than the Wherry (1952) or Landy and Farr (1980) models, and is meant only to aid here in forming hypotheses about possible approaches to successfully training raters.

possible, in other words, that in order to create positive changes in interrater agreement and accuracy of performance ratings, rather than (or probably in addition to) attending to rating errors in the training of raters, attention should be paid to training in behavior observation and to creating correct and well agreed upon standards of performance for raters to use in making judgments about ratee performance.

These speculations are in keeping with the Landy and Farr (1980) suggestion that raters develop common frames of reference for rating job performance and that they attend carefully to the performance requirements of the job. Other results from the literature that focus on the training of behavior observers generally support the promise of these approaches, as well (Jecker, Maccoby, & Brietrose, 1965; Spool, in press; Wahler & Leske, 1973). And, a study by Zuliani (personal communication) tentatively supports the notion of training raters to use standard (and correct) frames of reference. He found that military officers thoroughly trained in a particular leadership strategy provided leadership ratings that possessed high interrater agreement compared to the ratings made by persons not so trained. Research is needed to test these kinds of training approaches.

Overall, it appears that not enough is known about the effects of different types of training to make a judgment about its potential usefulness for enhancing rating accuracy. The Borman and Rosse study does suggest, however, that improvements in accuracy may be more difficult to bring about than simply changes in "rating behavior."

Administrative set as a contributor to rating accuracy. The most consistent finding here has been that a "for administration purposes" set (versus a "for research only" set) contributes to leniency error (e.g., Borreson, 1967; Taylor & Wherry, 1951). Of course, a perfectly consistent (across ratee) leniency error would not affect the differential accuracy (DA) of ratings, but severe leniency may well restrict the range of the ratings to the point where accuracy is adversely affected. Ratings in the military have been especially vulnerable to this phenomenon. Hollander (1957, 1965) has found that the interrater reliability and validity of ratings were not dependent on rating set (research vs. administrative), but clearly the severe leniency that may emerge in some situations is likely to reduce accuracy.

Conclusions. So what can we conclude about the effects of rater, format, training, and administrative set on rating accuracy? Just how important is each of these sources for helping raters evaluate others' performance more accurately?

A. First, it appears that rater individual differences are related substantially to rating accuracy. Evidence was presented suggesting a reasonable degree of stability in individuals' abilities to rate performance accurately, and 17% of the variance in rating accuracy can be accounted for by certain ability and personality variables. Of course, very different rating situations may alter these individual differences-rating accuracy relationships, but if stability emerges across a wide variety of performance rating situations, we might seriously consider selecting raters where possible.

Selecting raters according to their predicted "ability to rate" might be feasible in cases where many persons are available to provide peer ratings (such as in military basic training). It may, for example, be preferable to gather ratings from 10 highly qualified peers, peers with personal qualities associated with accurate ratings, than to gather ratings from 10 qualified and 5 to 10 relatively unqualified peers.

Selecting raters may also be realistic for jobs in which making performance appraisals is very important. For example, certain management jobs may require the incumbent to generate very precise and accurate performance assessments and perhaps selecting persons for these jobs at least partially on the basis of their predicted "ability to rate" would be a reasonable procedure to follow.

B. Common sense and a limited number of research studies suggest that "position to rate" is important, but the studies we reviewed yield no good estimate of "proportion of variance accounted for" by this variable. Nonetheless, the magnitude of the differences between the validities obtained by individuals in good position to rate versus those in poor position to rate in the Freeberg (1969) and Whitla and Tirrell (1953) studies indicates that the increment in accuracy can be considerable. It appears very important to select as raters only those individuals who have good opportunity to observe ratees' work behavior relevant to the dimensions being rated.

C. Rating format has proven to account for very little variance in performance rating accuracy. Some evidence: First, carefully developed formats, whether they be BARS, summated, Guttman type, etc., have typically been associated with approximately equal amounts of psychometric error. Second, Borman and Dunnette (1975) found that less than 5% variance in psychometric criteria was accounted for by differences in three formats (trait, BARS, and numerically anchored). And third, the Borman and Rosse (1978) study, providing a direct estimate of variance accounted for in accuracy by four different rating forms, revealed that 2% to 3% variation in accuracy was accounted for by differences between the formats. Thus, rating formats, at least the way we have been studying them, appear to have a relatively limited effect on performance rating accuracy.

D. As mentioned, we know very little about the effects of training raters on their subsequent performance rating accuracy. On the basis of the few studies done and my own experience in working with raters, I believe training may well prove to be very useful in enhancing the accuracy of ratings, but we don't presently know what kinds of training will have the desired positive effect.

E. No exact "variance accounted for" estimate can be made regarding the effects of administrative factors on performance rating acuracy; however, we should note that all of the above four approaches assume that the rater will try to make as accurate ratings as possible. Clearly, if the rater decides to distort his/her ratings, the result is utter disaster. In other words, if anything about the administrative set encourages raters to provide inaccurate ratings, all other efforts to encourage accurate ratings will necessarily be rendered insufficient. "Success" in getting the rater to at least try to record accurate performance judgments is a prerequisite for obtaining accurate ratings. Therefore, the importance of administrative set cannot be overemphasized.

2. Current Limitations in Performance Rating Accuracy and Possibilities for "Getting Beyond" Present Levels of Precision

In this section of the paper, I discuss possible limits in the precision with which performance ratings can be made and then examine one possible approach toward removing those limitations.

Possible limits in performance rating accuracy. Up to now my treatment of the various factors that affect accuracy have implicitly assumed that raters might attain significantly higher levels of accuracy than they do presently. Yet some recent research questions this assumption.

Kavanagh, MacKinney, and Wolins (1971) observed similar levels of convergent and discriminant validities for three studies involving performance ratings and concluded that ratings may suffer from a "characteristic error," limiting the precision with which raters make performance judgments. Picking up where Kavanagh et al. left off, I recently attempted to create relatively ideal conditions for obtaining performance ratings and then evaluated the precision of the resultant ratings (Borman, 1978). This study sought to establish approximate "ceiling benchmarks" for levels of interrater agreement and discriminant validity, essentially numerical indices of Kavanagh et al.'s characteristic error.

Accordingly, 14 "expert raters" carefully viewed the videotaped performances referred to previously, studied transcripts of the dialogue contained in the performances, and then used BARS rating forms to make their evaluations. Conditions for obtaining precise ratings were therefore quite favorable: raters were very knowledgeable about the two jobs depicted on the tapes and also about rating research and rater errors; they had ample opportunity to study the performances; and they used carefully developed rating forms in making their evaluations.

Convergent and discriminant validity results appear in Table 3. The intraclass indices obtained can be compared[3] to those found in other studies, these comparisons indicating that discriminant validity (especially) is substantially higher than it has been in other studies.

Nonetheless these ratings are far from perfect. For example, in four instances the most discrepant raters of the 14 disagreed in their ratings (of a particular ratee on a dimension) by five scale points (on a 1-7 scale). In 31 instances, four scale points separated the most discrepant raters. Now, it is true that when we focus on the overall stability of the ratings, a different picture emerges. Intraclass correlations for individual dimensions (measures of the reliability of the mean ratings) vary from .91 to .98 with a median of .97. But these summary indices of stability mask the sometimes glaring discrepancies in the expert judgments. Perfect interrater agreement in ratings does not guarantee perfect accuracy, but rater disagreement ensures the presence of rater error. Thus, it is clear that considerable imprecision remains in these ratings despite the care taken to create an ideal rating environment.

My conclusions from all this are, first, that there is plenty of room for improvement in the performance ratings we typically gather, even those gathered very carefully in research settings. A glance at the differences shown in Table 3 between the "ceiling benchmark" intraclass indices and those representing ratings gathered in the field strongly suggests that significant strides can be taken to improve (especially) the discriminant validity of performance ratings.

However, in my view, to get beyond the levels of rating precision gained in our study (Borman, 1978), we must examine the performance evaluation process to discover how perceptual and rating errors are made. Increased understanding of the performance evaluation process may suggest ways to reduce substantially the rating errors that limit performance rating accuracy, and perhaps we can then get beyond the "barriers" reflected in the Borman (1978) study.

Admittedly, I have no clear-cut plan pinpointing exactly how increased knowledge of the rating process will make possible higher accuracy in ratings. However, it seems to me that learning more about the rating

[3] Kavanagh et al. (1971) argue that these indices may be used to compare (across studies) degrees of convergent and discriminant validity and a kind of halo (rater x ratee interaction).

Table 3

Intraclass Indices for Seven Studies

	Intraclass Indices		
Study/Type of Raters	Ratee Effects	Ratee x Dimension Interaction	Rater x Ratee Interaction
Lawler (1967)			
Superior - peer	.63	.40	.31
Superior - self	.36	.19	.44
Kavanagh, MacKinney, & Wolins (1971)			
Superior - subordinate (all dimensions)	.44	.13	.50
Traits only	.56	.15	.62
Performance dimensions only	.44	.13	.50
Five selected dimensions only	.66	.17	.60
Borman, Toquam, & Rosse (1978)			
Self - peer - supervisor			
Traits	.45	.16	
Behavior scales	.42	.12	
Multidimensional scales	.37	.13	
Factor scales	.54	.11	
Boruch, Larkin, Wolins, & MacKinney (1970)			
Subordinates	.40	.20	.43
Schneier (1977)			
Peers (all dimensions)	.59	.16	.09
Five personal criteria	.69	.31	.21
Five performance criteria	.51	.08	.01
Dickinson & Tice (1973)			
Superiors and peers	.48	.13	.38
Borman (1978) - "Upper Limit" Estimates			
Psychologists and graduate students	.64	.57	.12
Psychologists and graduate students	.69	.58	.16

process is likely to suggest in turn ways of improving rating forms, redesigning rater training programs, selecting "qualified" raters, and forming guidelines for rater observational practices (in addition to the scientific contributions this kind of research would offer). For example, if we learned that most raters tend to use very little information (few dimensions) in making ratings but that a few very accurate raters use more information, we might attempt to train raters of the first type to take account of more information about ratees. Or, if we found that peers' evaluations or organizational skills were strongly influenced by their judgments of interpersonal skills, rendering the organizational skills ratings relatively inaccurate (relative to supervisor ratings), we might have supervisors provide the organizational skills rating. In other words, increasing our knowledge about elements of the rating process appears to represent a logical approach toward making more educated hypotheses about how to train, position, select, etc. raters. Optimistically, I contend that this strategy may even accomplish significant breakthroughs to get us beyond our present levels of precision in performance ratings. (More pessimistically, but still important, it may provide reasons why we can't expect to progress beyond certain levels of precision.) Thus, if this view is correct, we must learn considerably more about the performance rating process. Of course, articulating that view is much easier than acting on it. The next section discusses some ideas for studying rater process.

Current progress in and possibilities for studying the performance rating process. Social and personality psychologists have been studying the interpersonal perception process for many years, and industrial psychologists have much to learn about performance ratings from these efforts. The various approaches taken to examine facets of the person perception process are too many in number to review here. However, examples include studies of impression formation (Asch, 1946), an examination of implicit personality theories (Bruner & Tagiuri, 1954; Schneider, 1973), and investigations of various individual differences constructs linked to person perception-cognitive complexity (Bieri, 1961). Parallel efforts in the area of performance ratings definitely seem in order.

Also, "closer to home," the decision making process in the selection interview has been studied (Webster, 1964), and some of the results from that research may be useful to us. And, still another "source of inspiration" for rating process research may come from the extensive information processing and decision making research pioneered by Brunswik (e.g., Brunswik, 1955) and carried forward by Edwards (e.g., Edwards, 1954) and more recently Slovik and Lichtenstein (1971), along with many others. Man as information processor and decision maker has been characterized by both Bayesian and correlational models, and perhaps one or both of these models may contribute to a study of judgments about performance (e.g., Zedeck & Kafry, 1977).

Finally, a body of research that has some potential relevance for learning about rating process is the study of cognitive style. For example, a reliable distinction has been made between analyzers and synthesizers (Vernon, 1952), where the analytic observer concentrates on detail and may fail to integrate separate perceptions while the synthesizer is more likely to perceive things as an integrated whole at the expense of seeing some of the detail. Likewise, constructs such as flexibility of closure (Botzum, 1951), field dependence or psychological differentiation (Witkin, Dyk, Taterson, Goodenough, & Karp, 1962), and leveling/sharpening (Holzman & Klein, 1954) are perceptual individual differences that have been shown to correlate significantly with certain ability and personality variables, and they may also be associated with different approaches to rating performance.

With all of these potential sources for direction in studying the performance rating process, investigations into this process are only just beginning, and, as we shall see, most provide information that only indirectly adds to our knowledge of that process. One such indirect approach is to search for and investigate correlates of rating behavior. For example, one study (Crooks, 1972) found that Black and White supervisors gave slightly higher ratings to subordinates of their own race. Two other studies suggested that raters give higher evaluations to same sexed subordinates (Bigoness, 1976; Hamner, Kim, Baird, & Bigoness, 1974). And another study of this type found that the education and experience levels of police officer raters had little effect on their performance ratings (Cascio & Valenzi, 1977).

A closely related kind of study seeks other kinds of rater characteristics as correlates of rating behavior. I have already mentioned our study which examined ability, personality, vocational interest, and background correlates of rating accuracy (Borman, 1979). Mullins, Seidling, Wilbourn, and Earles (1979) conducted a somewhat similar study but found near zero correlations between rating accuracy and the individual differences they measured. Schneier (1977) found that cognitively complex raters tended to prefer a relatively complex BARS form and to make fewer rating errors on this scale, while cognitively "simple" raters preferred and made fewer such errors on a simpler rating format. However, a study by Bernardin and Boetcher (1978) failed to replicate this result.

To repeat, however, the studies briefly reviewed above focus only indirectly on the performance rating process. Where we are almost totally lacking is in studies that examine the rating process directly--by that I mean studies which investigate rating "style," different strategies for making ratings, preferences for using certain cues, methods of combining information to arrive at ratings, etc.

Two notable exceptions to this dearth of rating process research are studies by Zedeck and Kafry (1977) and Banks (in progress). I now briefly review these studies because they reflect the kinds of "rating process" research I have in mind. Zedeck and Kafry employed policy capturing methods to study the weights on different performance dimensions used by raters when making overall performance evaluations. They constructed 40 vignettes of hypothetical nurse ratees by including in a description of each nurse behavioral examples representing performance on each of the nine performance dimensions. Raters in the study evaluated the overall effectiveness of each "ratee," and the policy capturing analyses yielded the weights each rater used in making these evaluations. In addition, Zedeck and Kafry employed the Judgment Analysis technique (JAN; Christal, 1963) to cluster together raters with similar policies (i.e., similar patterns of weights), and then correlated several cognitive and vocational interest measures with cluster membership. Members of the two clusters revealed by JAN did not differ significantly on any of the individual differences measures. Despite the latter negative finding, this study represents an important attempt to discover how different raters use and combine information in making performance ratings.

The other research noted above (Banks, in progress) is a doctoral dissertation study by Cris Goggio Banks at the University of Minnesota. She has devised a procedure to identify behavioral cues raters use in making ratings. In particular, her study will examine: (a) the interrater agreement associated with use of these cues; (b) the kinds of cues raters use in making judgments on different types of performance dimensions; and (c) similarities and differences in the effectiveness levels attached to particular behaviors when they are used as cues for different dimensions.

In this study, Banks had each subject view one of the 5- to 9-minute manager performances on our videotapes and attend to the effectiveness displayed on a single dimension (different subjects were assigned different dimensions). Subjects had before them a computer console with seven buttons corresponding to seven effectiveness levels (1 = very ineffective; 7 = very effective), and they were instructed to press one of those buttons each time they viewed behavior they thought was relevant to the performance dimension being considered. The particular button pushed (of the seven) indicated the rater's judgment regarding the effectiveness of the behavior. Also, the buttons were attached to a timing device that provides an exact record of where in the tape each button was pressed. Finally, subjects were instructed to provide a brief verbal description of the behavior they were attending to each time they pressed a button.

In this manner, Banks intends to study aspects of the performance rating process very directly. This method may provide an effective vehicle for discovering at a micro-level some of "what is going on in raters' heads" as they go through the performance evaluation process. In the third and final part of this paper, I will describe an application of this method to studying accuracy and the rating process.

Conclusions. The central points made in this section are, first, that we now have some idea of the present level of precision we can expect from ratings when conditions are close to "ideal." It appears that the discriminant validities of field ratings might be improved considerably if conditions can be made more conducive to accurate ratings. Second, the extent of the various imprecisions and rater errors contained in our expert ratings was used to argue that new breakthroughs in the levels of performance rating accuracy await further understanding of the rating process. Past and current research on the interpersonal perception process may provide clues about how to proceed in studying the performance rating process. Also, research on information processing, decision making, and cognitive style should be attended to for possible applications related to studying performance ratings. Research on the performance rating process has only just begun and, sadly (in my view), most of that work only indirectly focuses on rating process. Two studies were then offered as examples of more direct assaults on learning about this process.

3. One Possible Approach Toward Exploring the Performance Rating Process

In this final part of the paper, I offer a strategy for studying the performance rating process and possible linkages between: (a) cognitive and personality individual differences; (b) a class of variables I will term "rating style" variables; and (c) halo and accuracy in performance ratings. I will force myself to be as explicit as possible about the proposed research strategy in reaction to a tendency on the part of industrial psychologists to talk about studying the performance rating process but then failing to specify how we might actually go about accomplishing this. Here is how the study might proceed.

First, the proposed research leans heavily on the innovative technology developed by Cris Goggio Banks. Also, it utilizes the manager videotapes we developed in a study done for the U. S. Army Research Institute (Borman, Hough, & Dunnette, 1976).

An early step in the study would be to examine transcripts of Banks' thousands of "behavioral cue" responses and develop one or more category systems that describe the kinds of cues raters use in making performance judgments. (Recall that Banks asked her subjects to report each cue they used; i.e., the behaviors or other kinds of cues they saw as important indicators of performance on the target dimension.) The category systems envisioned would be the result of a content analysis of these many transcribed reports. Possible categories that might emerge, for example, are interpersonal-related vs. thing-related, directly behavioral vs. inferential, and something the ratee did vs. something another person did in reaction to the ratee. The categories would be used subsequently to classify the same kinds of responses made by members of another sample.

The next step would require that a "new" sample be administered a battery of ability, personality, vocational interest, and background inventories, probably similar to the battery we developed for a previous study (Borman et al., 1976). Then these subjects would perform much the same rating task as Banks' subjects did. Specifically, we would select from the manager job two dimensions that represent the kinds of dimensions relevant to a variety of jobs. This is because the study's results are more likely to have meaning in other settings if these dimensions are "representative." Subjects would then view each of the eight manager tapes, perform the button-pressing task (reporting the cues they were attending to and the effectiveness of the behavior noted), and also rate each performer's effectiveness on the target dimension. Then some time later the subjects would go through the same procedure again, this time focusing on the other dimension.

Now, researchers employing the categories developed previously would help each subject sort his/her transcribed "cue reports" into the proper categories, and the reports for each subject would also be "scored" in a number of other ways in order to further describe the rating "style" utilized by the subject. Some examples: (a) total number of button presses would provide a "raw" measure of the number of cues entering into a subject's ratings; (b) average variance of the effectiveness levels associated with the button presses (for individual ratees) would yield a measure of the tendency to notice (and gather) both favorable and unfavorable information related to individuals' performances; and (c) comparisons of the mean cue effectiveness ratings (associated with individual ratees) to the corresponding dimension rating made by the subject might provide a score indicating a tendency to overweight bad (or good) information.

We envision scoring these data many ways in an effort to characterize various individual differences in rating style.[4] The resultant scores for each subject rater can then of course be correlated with the cognitive, personality, etc., individual differences measures and with accuracy (DA) and halo. As mentioned, this matrix of correlations should provide important clues about the links between these variable sets.

One last way to view the contribution of such a study is to examine what it might tell us about each of the three elements in Borman's (1978) simple three-step model of the performance rating process. First, the content of the reported cues would provide useful, basic information about exactly what it is that raters attend to when making

[4] Notice that the data are configured in such a way that the intra-subject stability of these scores can be evaluated. This is a very important first step in assessing the usefulness of these individual differences for describing rating style.

performance judgments (Step a). Second, the effectiveness levels assigned by subjects to each cue, including tne patterns of these effectiveness levels, should help us gain some understanding of Step b of the model. And third, relationships between the effectiveness levels a rater assigns to the various cues associated with a ratee's performance and the "summary" rating the rater assigns to the ratee on a dimension should contribute to our understanding of Step c.

This reasonable explicit example of possible research exploring the performance rating process should serve to demonstrate one type of research project that can be done in this area. And, as has been argued in this paper, greater understanding of the performance rating process may be necessary if we are to experience significant progress in increasing performance rating accuracy.

References

Adams, H.F. The good judge of personality. Journal of Abnormal and Social Psychology, 1927, 22, 172-181.

Amir, Y., Kovarsky, Y., & Sharan, S. Peer nominations as a predictor of multistage promotions in a ramified organization. Journal of Applied Psychology, 1970, 54, 462-469.

Asch, S.E. Forming impressions of personality. Journal of Abnormal and Social Psychology, 1946, 41, 258-290.

Banks, C.G. Analyzing the rating process: A content analysis approach. Unpublished doctoral dissertation, University of Minnesota, in progress.

Bernardin, H.J. Behavioral expectation scales versus summated scales--A fairer comparison. Journal of Applied Psychology, 1977, 62, 422-427.

Bernardin, H.J. Effects of rater training on leniency and halo errors in student ratings of instructors. Journal of Applied Psychology, 1978, 63, 301-308.

Bernardin, H.J., Alvares, K.M., & Cranny, C.J. A recomparison of behavioral expectation scales to summated scales. Journal of Applied Psychology, 1976, 61, 564-570.

Bernardin, HM.J., & Boetcher, R. The effects of rater training and cognitive complexity on psychometric error in ratings. Paper presented at the American Psychological Association Convention, 1978.

Bernardin, J.H., & Walter, C.S. Effects of rater training and diary keeping on psychometric error in ratings. Journal of Applied Psychology, 1977, 62, 64-69.

Bieri, J. Cognitive complexity-simplicity as a personality variable in cognitive and preferential behavior. In D.W. Fiske & S.R. Maddi (Eds.), Functions of varied experience. Homewood, IL: Dorsey, 1961.

Bigoness, W.J. Effect of applicant's sex, race, and performance on employers' performance ratings: Some additional findings. Journal of Applied Psychology, 1976, 61, 80-84.

Borman, W.C. The rating of individuals in organizations: An alternate approach. Organizational Behavior and Human Performance, 1974, 12, 105-124.

Borman, W.C. Effects of instructions to avoid halo error on reliability and validity of performance evaluation ratings. Journal of Applied Psychology, 1975, 60, 556-560.

Borman, W.C. Consistency of rating accuracy and rating errors in the judgment of human performance. Organizational Behavior and Human Performance, 1977, 20, 258-272.

Borman, W.C. Exploring upper limits of reliability and validity in job performance ratings. Journal of Applied Psychology, 1978, 63, 135-144.

Borman, W.C. Individual differences correlates of accuracy in evaluating others' performance effectiveness. Applied Psychological Measurement, 1979, 3, 103-115.

Borman, W.C., & Dunnette, M.D. Behavior based vs. trait oriented performance ratings: An empirical study. Journal of Applied Psychology, 1975, 60, 561-565.

Borman, W.C., Hough, L.M., & Dunnette, M.D. Performance ratings: An investigation of reliability, accuracy, and relationships between individual differences and rater error. Minneapolis: Personnel Decisions, Inc., 1976.

Borman, W.C., & Rosse, R.L. Format and training effects on rating accuracy and rater errors. Minneapolis: Personnel Decisions Research Institute, 1978.

Borman, W.C., Toquam, J.L., & Rosse, R.L. Development and validation of an inventory battery to predict Navy and Marine Corps recruiter performance. Minneapolis: Personnel Decisions Research Institute, 1978.

Borreson, H.A. The effects of instructions and item content on three types of ratings. *Educational and Psychological Measurement*, 1967, *27*, 855-862.

Boruch, R.F., Larkin, J.D., Wolins, L., & MacKinney, A.C. Alternative methods of analysis: Multitrait-multimethod data. *Educational and Psychological Measurement*, 1970, *30*, 833-853.

Botzum, W.A. A factorial study of the reasoning and closure factors. *Psychometrika*, 1951, *16*, 361-386.

Brown, E.M. Influence of training, method and relationship on the halo effect. *Journal of Applied Psychology*, 1968, *52*, 195-199.

Bruner, J.S., & Tagiuri, R. The perception of people. In G. Lindzey (Ed.), *Handbook of social psychology*. Reading, Mass.: Addison-Wesley, 1954.

Brunswik, E. Representative design and probabilistic theory in a functional psychology. *Psychological Review*, 1955, *62*, 193-217.

Buckner, D.N. The predictability of ratings as a function of interrater agreement. *Journal of Applied Psychology*, 1959, *43*, 60-64.

Campbell, J.P., Dunnette, M.D., Lawler, E.E., & Weick, K.E. *Managerial behavior, performance, and effectiveness*. New York: McGraw-Hill, 1970.

Cascio, W.F., & Valenzi, E.R. Behaviorally anchored rating scales: Effects of education and job experience of raters and ratees. *Journal of Applied Psychology*, 1977, *62*, 278-282.

Christal, R.E. *JAN: A technique for analyzing group judgment*. PRL-TDR-63-3, AD 403 813, Lackland AFB, TX: Personnel Research Laboratory, Aerospace Medical Division, February 1963.

Cline, V.A. Interpersonal perception. In B.A. Maher (Ed.), *Progress in experimental personality research* (Vol. 1). New York: Academic Press, 1964, 221-284.

Cronbach, L.J. Processes affecting scores on understanding of others and "assumed similarity." *Psychological Bulletin*, 1955, *52*, 177-193.

Crooks, L.A. (Ed.). *An investigation of sources of bias in the prediction of job performance: A six year study*. Princeton, NJ: Educational Testing Service, 1972.

Crow, W.J. The effect of training upon accuracy and variability in interpersonal perception. *Journal of Abnormal and Social Psychology*, 1957, *55*, 355-359.

Crow, W.J., & Hammond, K.R. The generality of accuracy and response sets in interpersonal perception. Journal of Abnormal and Social Psychology, 1957, 54, 384-390.

Dickinson, T.L., & Tice, T.E. A multitrait-multimethod analysis of scales developed by retranslation. Organizational Behavior and Human Performance, 1973, 9, 421-438.

Dunnette, M.D. Forms of interpersonal accommodation: Processes, problems, and research avenues. Paper presented at the American Psychological Association Convention, San Francisco, September 1968.

Dunnette, M.D., & Borman, W.C. Personnel selection and classification systems. Annual Review of Psychology, 1979, 30, 477-525.

Edwards, W. The theory of decision making. Psychological Bulletin, 1954, 51, 380-418.

Edwards, B.C., & McWilliams, J.M. Expressor sex, perceiver personality, and cognitive perception. Journal of Psychology, 1974, 87, 137-141.

Einhorn, H.J. Expert measurement and mechanical combination. Organizational Behavior and Human Performance, 1972, 7, 86-106.

Estes, S.G. Judging personality from expressive behavior. Journal of Abnormal Psychology, 1938, 33, 217-236.

Freeberg, N.E. Relevance of rater-ratee acquaintance in the validity and reliability of ratings. Journal of Applied Psychology, 1969, 53, 518-524.

Gage, N.L., & Cronbach, L.J. Conceptual and methodological problems in interpersonal perception. Psychological Review, 1955, 62, 411-422.

Green, G.H. Insight and group adjustment. Journal of Abnormal and Social Psychology, 1948, 43, 49-61.

Hamner, W.C., Kim, J.S., Baird, L., & Bigoness, W.J. Race and sex as determinants of ratings by potential employers in a simulated work-sampling task. Journal of Applied Psychology, 1974, 59, 705-711.

Hastorf, A.H., Schneider, D.J., & Polefka, J. Person perception. Reading, MA: Addison-Wesley, 1970.

Hjelle, L.H. Personality characteristics associated with interpersonal perception accuracy. Journal of Counseling Psychology, 1969, 16, 579-581.

Hollander, E.P. The reliability of peer nominations under various conditions of administration. Journal of Applied Psychology, 1957, 41, 85-90.

Hollander, E.P. Validity of peer nominations in predicting a distant performance criterion. Journal of Applied Psychology, 1965, 49, 434-438.

Holzman, P.S., & Klein, G.S. Cognitive system-principles of leveling and sharpening: Individual differences in visual time-error assimilation effects. Journal of Psychology, 1954, 37, 105-122.

Jecker, J.O., Maccoby, N., & Brietrose, H.S. Improving accuracy in interpreting cues of comprehension. Psychology in the Schools, 1965, 2, 239-244.

Kavanagh, M.J., MacKinney, A.C., & Wolins, L. Issues in managerial performance: Multitrait-multimethod analysis of ratings. Psychological Bulletin, 1971, 75, 34-49.

Landy, F.J., & Farr, J.L. Performance rating. Psychological Bulletin, 1980, 87, 72-107.

Latham, G.P., Wexley, K.N., & Pursell, E.D. Training managers to minimize rating errors in the observation of behavior. Journal of Applied Psychology, 1975, 60, 550-555.

Lawler, E.E. The multitrait-multirater approach to measuring managerial job performance. Journal of Applied Psychology, 1967, 51, 369-381.

Levine, J., & Butler, J. Lecture versus group discussion in changing behavior. Journal of Applied Psychology, 1952, 36, 29-33.

Mullins, C.J., & Force, R.C. Rater accuracy as a generalized ability. Journal of Applied Psychology, 1962, 46, 191-193.

Mullins, C.J., Seidling, K., Wilbourn, J.M., & Earles, J.A. Rater accuracy study. AFHRL-TR-78-89, AD A066 779. Brooks Air Force Base, TX: Personnel Research Division, Air Force Human Resources Laboratory, February 1979.

Schneider, D.J. Implicit personality theory. Psychological Bulletin, 1973, 79, 294-309.

Schneier, C.E. Operational utility and psychometric characteristics of behavioral expectation scales: A cognitive reinterpretation. Journal of Applied Psychology, 1977, 62, 541-548.

Schwab, D.P., Heneman, H.G., & DeCotiis, T. Behaviorally anchored rating scales: A review of the literature. Personnel Psychology, 1975, 28, 549-562.

Scodell, A., & Mussen, P. Social perceptions of authoritarians and nonauthoritarians. *Journal of Abnormal and Social Psychology*, 1953, 43, 181-184.

Sechrest, L., & Jackson, D.N. Social intelligence and accuracy of interpersonal predictions. *Journal of Personality*, 1961, 29, 167-182.

Slovik, P., & Lichtenstein, S. Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance*. 1971, 6, 649-744.

Spool, M.D. Training programs for observers of behavior: A review. *Personnel Psychology*, in press.

Taft, R. The ability to judge people. *Psychological Bulletin*, 1955, 52, 1-23.

Taylor, E.K., & Hastman, R. Relation of format and administration to the characteristics of graphic rating scales. *Personnel Psychology*, 1956, 9, 181-206.

Taylor, E.K., & Wherry, R.J. A study of leniency in two rating systems. *Personnel Psychology*, 1951, 4, 39-47.

Tupes, E.C. *Relationships between ratings by peers and later performance of USAF Officer Candidate School graduates*. (AFPTRC-TN-57-124, ASTIA Document No. AD-134 257). Lackland Air Force Base, TX: Air Force Personnel and Training Research Center, October 1957.

Tupes, E.C. *Personality trait related to effectiveness of junior and senior Air Force officers* (WADC-TN-59-198). Lackland Air Force Base, TX: Personnel Laboratory, Wright Air Development Command, November 1959.

Vernon, M.D. *A further study of visual perception*. Cambridge, MA: Cambridge University Press, 1952.

Vernon, P.E. Some characteristics of the good judge of personality. *Journal of Social Psychology*, 1933, 4, 42-57.

Vingoe, F.J., & Antonoff, S.R. Personality characteristics of good judges of others. *Journal of Counseling Psychology*, 1968, 15, 91-93.

Wahler, R.G., & Leske, G. Accurate and inaccurate observer summary reports. *Journal of Nervous and Mental Disease*, 1973, 156, 386-394.

Waters, L.K., & Waters, C.W. Peer nominations as predictors of short-term sales performance. *Journal of Applied Psychology*, 1970, 54, 42-44.

Webster, E.C. Decision making in the selection interview. Montreal: Eagle, 1964.

Wedeck, J. The relationship between personality and "psychological ability." British Journal of Psychology, 1947, 37, 133-151.

Wherry, R.J. The control of bias in rating: A theory of rating. PRB Report No. 922, Contract No. DA-49-083 OSA 69, Department of the Army, 1952.

Whitla, D.K., & Tirrell, J.E. The validity of ratings of several levels of supervisors. Personnel Psychology, 1953, 6, 461-466.

Witkin, H.A., Dyk, R.B., Taterson, H.F., Goodenough, D.R., & Karp, S.A. Psychological differentiation. New York, NY: Wiley, 1962.

Zedeck, S., & Kafry, D. Capturing rater policies for processing evaluation data. Organizational Behavior and Human Performance, 1977, 18, 269-294.

Zuliani, R. Personal communication to author, 1978.

Dr. Borman: We'll probably end up trying a number of different formats. The format I'd like to try is giving you a brief summary of my paper and then opening it to questions. I also have a list of issues to talk about in case no discussion comes forth. I don't think we'll have any problem with that but I came super-prepared just in case.

An early point I raised in my paper was that, my feeling at least, performance rater accuracy or validity is pretty clearly the most important psychometric consideration. We are forced, for obvious reasons normally, not to look at accuracy or validity but instead such other kinds of psychometric considerations as leniency, halo, restriction of range, perhaps interrater agreement. But any time we have a situation--and this is normally the case in any kind of field study--where there is no kind of criterion true score that you can attach to the performance of an individual, there's really no way--unless we have some kind of super, external criterion that everybody agrees is an objective, external kind of criterion--you really have no way to assess validity. So the attention paid to these other kinds of measures is quite understandable.

The reason I'm a bit concerned about this recently though is that it seems to me that the relationship between other kinds of psychometric criteria and accuracy or validity may not be as high as we had assumed. At least I had assumed. I had assumed that if you reduce the leniency in a set of ratings or reduce the halo in a set of ratings, obviously you have better ratings and probably more accurate ratings.

Some recent data that we have gotten recently, and some data that we've gotten out of other studies, have shown at least indirectly that these relationships may not be very high, which is very upsetting because when we're studying leniency and restriction of range and other kinds of psychometric criteria, we may not have been doing anything to accuracy. But I reviewed various attempts either direct or indirect to increase rating accuracy. In an attempt to make some kind of proportion-of-variance-accounted-for statement, that is, proportion-of-variance-accounted-for in increases in rating accuracy by these various kinds of interventions or sources, for the most part I could not make exact estimates. But I still tried.

The first kind of source that I looked at was rater individual differences, and some data that we have show that perhaps as much as 15% to 20% of the variance in performing rating accuracy may be due to rater individual differences; such things as intelligence, detail orientation, personal adjustment. I mean this is quite shocking, to me at least. I didn't really expect this. It may well be that what this suggests is that, where we can, we should perhaps be selecting raters to make performance evaluations, and I came up with a couple of examples where you might, practically speaking, actually be able to pull this off.

In a peer rating situation where you have many, many peers you may be able to gather data from only some of the peers, rather than all of the peers--those that are the most highly qualified to make performance evaluations based on the reading of their individual differences, of their standing in individual differences.

Also, I looked quickly at rater position, that is organizational level, for instance, that the rater is assigned to, relation to the ratee, and could not really come up with any kind of proportion-of-variance-accounted-for statement, but concluded that it seems to be a very important kind of factor, especially if you get into rather bizarre kinds of rating assignments where you have a supervisor two or three levels removed making a performance rating. Obviously, common sense tells us that's just stupid because the ratee behavior observed is just not great enough or relevant enough to make any kind of a sensible evaluation from that kind of standpoint. So it's something I concluded that we should attend to, but common sense pretty well takes over.

Rater training as another kind of source--I concluded that we just don't know enough about how to improve rater accuracy through training to make any kind of an absolute statement about this. Some people in this room are moving forward in that kind of area and I think that we may know more about this soon, but my gut feeling in working with raters both in the field and in the laboratory is that training has to do some good. I've run into quite a number, a disgustingly large number of raters who, when you talk to them about what they're about to do, you know they really have the wrong idea, they have a very distorted view about how they're to go about rating performance. Obviously, if you give them a little training, it may be very simple, just a little talking to, to straighten them out as to what they're supposed to be doing, it's got to make a difference, but we haven't really gotten at the exact elements of rater training that might be useful in this regard.

One fairly solid conclusion I was able to arrive at was that rating format now appears to have a fairly limited kind of impact on rating accuracy. I would just say, throwing out numbers, that perhaps 2% to 5% of the variance in performance rating accuracy may be attributable to rating format kinds of design. Again, this gets thrown way off if we're talking about designing a form that's absolutely ridiculous, one which confuses the raters; e.g., instructions that are impossible to read. Obviously in that kind of situation rating format could control more of the variance in performance rating accuracy. But for well-developed formats, I think we're talking about really a fairly small proportion of variance accounted for.

Finally, I did something very limited in the administrative set area, mainly because I don't know that much about it, I guess. I think we're going to be tackling that problem in the conference here and it's obviously a very important kind of source of rating variance. I kind of concluded that it seems to me to be a prerequisite for gaining accurate ratings. I mean you've got to set the rater up so that he or she wants to make accurate ratings. Otherwise all these other things that we're talking about are just ridiculous. It doesn't matter what kind of training program you have or what kind of format you put in front of a person, you have to somehow set up the system so the rater at least wants to or is going to make an honest effort to provide accurate ratings.

The second part of the paper then went on to introduce another idea that all these kinds of approaches assume that we're going to be gaining someday much more accurate ratings, and some recent thinking and research at least questions this assumption. Mike Kavanagh's introduction of the so-called characteristic error which goes something like this, as I understand it, that there may be a certain limit beyond which people just could not make any more accurate or more precise ratings. Something in our cognitive structure may limit the precision with which we can make these kinds of ratings. We followed up on that in an empirical study and at least developed some kind of perhaps reasonable actual numerical limits in terms of conversion of discriminant validity especially using Mike's indices that he and MacKinney and Wolins introduced a while back. I actually used this research and this thinking to argue two different ways in a sense. On the one hand it seems to me that, comparing the indices that we obtained in our study with the general level of convergent and discriminant validity that you obtain in the field, there is room for improvement, especially in the discriminant validity area or the degree to which people can reliably differentiate an individual's performance across a number of dimensions. I think there is room for improvement and that we at least have hope in that regard.

Then I used the same data to argue exactly the opposite way that, on the other hand, to get beyond the levels of precision that we found in our study (just a quick editorial about that), we looked at the data that we gathered from 14 very highly qualified raters and found just unbelievable disagreement under a very controlled situation in which they were viewing video tapes. They saw exactly the same behavior, they were using beautiful rating scales, and we still found unbelievable disagreement in certain cases between these fine, fine upstanding, expert compulsive raters. This is very discouraging because it means to me that we may have to learn much more about the rating process, about what's really going on in the raters' heads, how rating errors are actually made, to get beyond the kinds of levels of precision that we gained even with this fine upstanding group of raters.

I then made some observations in the paper that the current research in ratings focuses only indirectly on the rating process and went through a number of examples. It's not that it's bad research;

it's just that it's not getting at quite the thing that I'm thinking might be important.

Then I offered two studies that I think do directly focus on the rating process, a study by Zedeck and Kafry that's described in the paper and a study that you're undoubtedly not familiar with, by a graduate student at the University of Minnesota, Cris Banks, who is doing some very basic research into the performance rating process, such as what people really do cue on when they're trying to make performance evaluations. So I then also offered a number of other sources that we might turn to, to attack this rating ·ocess question, such research as the interpersonal perception research. Personality and social psychologists have been studying process for years, and a couple of people I am familiar with in this room have also looked into this literature quite a bit. But it's fairly rare for industrial psychologists to know anything about this, and it seems like a good area to give us some hints on.

Also closer to home, the selection interview process research, which has been popular at least since the early 70's, seems to be another reasonable place to look for hints about how to study the rating process.

Also, this is a little flaky perhaps, but the cognitive style research, looking at such things as field dependence, leveling versus sharpening, for some reason that all went out in the middle 50's. So I'm sure I'm way behind the time as far as being interested in this at all, but it seems to me that those kinds of cognitive styles may also be important for determining how accurately people can make ratings and it may be important to look at.

I have some issues that I'd like to talk about if other people don't start out.

Dr. Muckler: We're doing a series of studies on the comparative performance of males, females, and minority group members in non-traditional Navy jobs. Our suspicion is that no matter what we come up with somebody's going to be angry about it. We were surprised to find that there are a lot of women doing non-traditional Navy· jobs. I don't know how they got there because they're not supposed to be there, but nevertheless they are, so it's been our opportunity to go out and measure them in the field actually working. Now we're using a lot of objective performance measurements; we're also using a lot of rating techniques, both for supervisory and for peer appraisal, and of course we've got the problem of rater bias. We're using a lot of unusual techniques to try to identify what kind of biases people have when they make ratings. And so far we've been quite successful in identifying the kinds of people with very strong biases and what kind of biases they have. For example, we find that people who are biased strongly against women are also very strongly biased against minority groups. The biased raters that we find are biased, period. They just don't like anybody except certain kinds of people. We also find that if you're a little careful about checking rater bias the people pop up pretty quick. It's not too difficult to find out what kind of biases they have. And finally, if you exclude those people from the ratings, the validity coefficients just zoom upward, and so we've been . . . .

Dr. Borman: How do you identify them? Look at their mean ratings?

Dr. Muckler: We've got some tests like yours, Wally. We've got some video tape situations that we set up. It's standard tasks and we know exactly how well they're being performed and we have them performed by males and females, Black and White, Hispanic and White, although of course nothing is made of this. I mean they come in and just perform the tasks, and bias will really stand out with this kind of situation.

Dr. Bernardin: Well, that's amazing because the literature doesn't support that approach and if you can get biases in that artificial situation, you can imagine what kind of biases you get in the real world.

Dr. Cascio: When you say validity coefficient, is the criterion rater accuracy?

Dr. Muckler: No. Here we're going against actual job performance tasks that we have, so we're going out and watching them actually perform, which turns out, of course, not to be the easiest thing that you can do in our kind of situation. I'm very fond of the peer appraisal system because I think those are extremely effective provided you do identify that small percent of people who really aren't very objective about what they're rating.

Dr. Borman: I suggest that possibly we could select raters on the basis of their standing on individual differences that were correlated with accuracy. It [illegible] much more direct to actually have them perform some practice [illegible]ings like this, identify people in that way who are not going to [illegible] it right, and withdraw them. It seems like a much more direct approach than the indirect approach.

Dr. Cascio: To follow up on that, one of the things that occurred to me, as you were talking about looking into some of the research on the decision making process in the interview, and how cues are picked up, was the well-known LIAMA research, Carlson et al. One of those studies that really stands out in my mind is when they presented video taped selection interviews and gave raters paper and pencil to sit back and watch the interview. It was a 20-minute video tape. At the end of the 20 minutes, they asked them a series of questions about what they saw in the video tape, and, to get to the bottom line in a hurry, what they found was that those people who, as they were watching the video tape, wrote down what they were seeing, made notes, memory joggers, etc., were fantastically more accurate in recalling exactly what took place than those people who trusted their own memory. Of course Carlson concluded that those more accurate raters used an individual differences strategy, whereas those who used the global approach were much less accurate, and it sort of occurred to me right away in selecting raters, this attention to detail might be important.

Dr. Mullins: I have a couple of points that occurred to me, Wally. One of these has to do with the selection of raters. In our particular situation, we have two different problems. One of these is establishing some criterion against which we want to do research which of course is the criterion that has to be solved. In a situation like that we can use peer ratings. And I think this device of eliminating those with the most obvious biases appears to be an easier way to go than selecting one or two who are more accurate, and very likely it would work out to be the same thing in the long run. But when you get into the operational situation, it would be well nigh impossible to use peer ratings really as an evaluation because if anything important is hinging on it, you're going to have some "dog eat dog" effect entering into that and I doubt that the powers that be would go for having peer ratings as a standard way of establishing something of this type. So that leaves us with a situation of using supervisor ratings and still somehow selecting people who will rate more accurately. This also had occurred to me somewhat earlier and I ever had such crazy ideas as having a team go around from unit to unit composed of good raters, all of them on a supervisory level, watching the guy work for a day or two, which, as I say, is a little crazy, but other than doing something like that I don't see . . . . We're going to look into one way, possibly, of handling that problem, and that is, we're going to see if we can't weight the ratings of raters when they've already been appraised on their accuracy in rating to see if this will drive up the predictive possibilities of what we're trying to do. But that's down the road a piece yet.

Dr. Ree: We would run into an operational problem, though. If we start pulling regulars out of their supervisory positions or out of the positions that they hold, there are jobs where raters coming in for a short period of time may have an inadequate exposure to what is going on.

What perhaps we need, what I hope to propose later today, is that we need a way to put all the raters on the same scale, or to determine that the raters in fact are all rating the same thing. We know in the trivial case that if one rater rates everybody high and another rater rates everybody low, these people are obviously not rating the same thing as long as we assume that there is some difference.

I participated in a study about 1971, in the Philadelphia Naval Shipyard, which was a validity study based on work sampling techniques. We were looking at the possibility of using work sampling techniques for the selecting of lead machinists. These are the people who machine propellers and things like this. One of the things we did was collect a work sample actually to use as a predictor, but to be used in a sort of concurrent validity sense, and of course a work sample like that could just as well be a criterion measure if it's appropriately designed. We worked in a large shop with a number of supervisors and a number of machinists. One of the things we found was that we could get only very, very poor discrimination among the middle group, although the two ends were easy to make discriminations among. We had multiple supervisors rating these people, and all the supervisors could easily point out the failure, and the excellent individual.

So I have some trouble from this particular viewpoint as to "the validity of ratings." We look at these psychometric considerations--I don't know what they mean to us. Why should, for example, a graduate school like Princeton University expect a great deal of variability in the math ability of the people they select into their math graduate program? I should expect just the opposite. So variability and skew and the various moments of the distribution may be totally inaccurate and totally worthless for the type of thing we're doing.

The concept of working with these simulated ratings is very good. One has to know then, how does it carry over out into the field? These are some of the problems. I want to address some of these things later.

Dr. Mullins: The second thing that I wanted to mention is that, as you know, I wholeheartedly support the idea of getting some kind of validity for ratings whenever one is doing a rating study. I don't like the idea of appealing to internal psychometric characteristics of your set of ratings as reality. I realize it's sometimes very difficult to do that. You certainly can't build a new film for every situation that you go into, but we have developed a couple of very simple, quick methods that could be used as a substitute for the more carefully done things.

One of these is a technique of simply administering some test which everyone is familiar with, like a vocabulary test, and then asking the people to estimate the score made on that vocabulary test.

Now that is a perfectly definable, scorable, observable task, and then using that, of course, as a criterion.

Another way, and this is sort of an intermediate kind of criterion that we've used on a few studies, is to collect rating profiles on people and then give these profiles without any identifiers back to the group who did the ratings and ask them to indicate who these profiles came from. The reasoning is that if they can't even recognize the profiles of the people in the class or the group, it seems unlikely that the ratings are doing anything very descriptive of the person. If the ratings are not descriptive of the person, it seems unlikely that there can be very much usefulness in them. I've worded this in an extreme but of course you also have degrees of that. There will be people who identify other people very well, and there will be some situations where the ratings are simply so bad that nobody identifies nobody, except by chance. And other situations in which there are people who can't recognize any one of the profiles, whereas everybody else can.

So I would very strongly recommend that everytime we do a rating study, we throw in something to get some kind of external criterion so that you can check the accuracy, even if it's just a little Mickey Mouse thing like those two techniques.

Dr. Bernardin: I think I'll interject some comment to that, although they were in response to some of Wally's comments in his paper, but I recommend two things that are close to that. One is that while leniency is not a good measure of error and apparently is not correlated with accuracy in our lab studies, I found that most of those studies which do use accuracy as a dependent variable are training studies, almost all of them. Aren't they, Wally?

Dr. Borman: Training studies in that raters are trained?

Dr. Bernardin: Yes. I want to get into that in my paper about the problem with the response set in training and what may cause that breakdown there, whereas in the field when we're using leniency error, it may be related to validity, but we can't assess validity.

What you could do, if you're looking at leniency, if there are any objective measures that are available--one obvious one is absences--you have the raters rate on that one variable as well. If you've got the objective data as an external criterion, you can go back and relate those to one another and you could then infer that, if people are lenient on that particular aspect of performance, they may be--and this is questionable of course--they may also be lenient on other dimensions of performance. That's one thing I recommend.

Another thing which gets very close to what Cecil just recommended is I think that you should validate raters, not only rating scales or rating procedures. That could be done by having raters, on maybe a yearly basis or every couple of years, generate a list of critical incidents similar to your profile type of thing that would be a justification of a number on a person, if you're going to use numerical ratings. Then you take those incidents and you make them anonymous, meaning not referring to any particular ratee, and you go to another pool of raters and have them rate the effectiveness and importance of those critical incidents which can then be transferred back into a rating on each ratee. Then you can correlate the two together--the actual numerical rating assigned by the rater and the ratings of the critical incidents that were assigned by other raters who did not know that they were attached to particular ratees. This is another form of what I would consider validity when you don't have accuracy scores through tapes and such. I think that's two approaches to external validity that might work.

Dr. Borman: The evidence that I have indicates that the relationships between these different psychometric criteria are really flaky presently. Obviously, in the most severe cases, severe leniency will wipe out accuracy totally, if we're talking about a given rater rating everyone at the 9-level on a 9-point scale. So obviously there is an almost perfect relationship if we're talking about that kind of thing, but as to how these relate in a number of different rating situations, I would not be prepared to say. I'm just saying that it may be that they're not as highly correlated as at least we had thought in our own minds that they were correlated.

Dr. Kavanagh: I think that Wally's way of operationalizing the validity of the rating system is quite good. Primarily, all we have now is the multitrait-multimethod approach, which has had some difficulties. It's really a very good way of summarizing data, but there are other approaches to the validity question. It's useful, I think, in the way Wally's used it and in some other ways. That's important, and what John is saying is also important. I would like to add, John, that you have more than one objective measure. In addition to absenteeism, you might have about three or four others, depending on the job. But my question is a little different. In the field, do raters want to be more accurate?

Dr. Borman: I agree. That is a weird way of looking at it but it seems to me that if we are going to make some kinds of important personnel decisions based on ratings, they should be as accurate as possible. But I've also thought, as one of the papers states, that it may be more important to increase organizational effectiveness and forget about accuracy. In other words, for example, if it turns out that by rating everyone high, that is if all raters rate everyone high and this makes everybody feel good and perform more effectively than if there were this great fine differentiation of individuals on these sophisticated rating scales, then it really does form an argument for not looking so much at accuracy but at organizational effectiveness.

Dr. Kavanagh: One of the things we've been examining with a number of our management groups in class exercises is the confidence that an individual has in his or her rating of another individual. We feel this is an important independent variable to impact on that process of feedback in the appraisal interview. I was wondering when you were talking--if I can convince a rater that he or she will be more accurate, I may increase the confidence that they feel in their ratings, thereby increasing the feedback to the individual. That would be a payoff, and we've been examining it, and it is in some of my papers which I'll discuss later.

Dr. Mullins: I think in our situation you don't run across that problem nearly as often as you do the opposite one, and that is that raters feel they know good and well what reality is and they'd rather not have us mucking around in it. That happens to us now and then.

Dr. Kavanagh: That's an interesting thought. The reason I got onto this topic was because, while talking to a group of managers, I posed the question of increasing the quality of the feedback process. Their responses indicated they do lack confidence in their ratings. They are often confronted by the ratee, who says: "That's bull. That's not where I'm at on this job." I can see the opposite thing occurring too in an organization that is highly autocratic--and I'm not addressing this to your organization--but in highly autocratic organizations I can envision that people would say, "This is the way it is." Now I don't think we have any organizations that rule-bound around, with a few obvious exceptions. So I'm surprised to hear you say that. Do you mean that your ratees are never challenging their ratings?

Dr. Mullins: There are two systems that we have to talk about. Under the old system, nobody ever challenged it because it didn't make any difference. It was just a number that you went through every year, just an exercise, and that was it. On a scale of 100 we were averaging about 98-point-something. But who cares, it didn't make any difference anyway; it had nothing to do with anything. Under the new system the preliminary tryouts that we've run have indicated that both supervisors and supervisees like the new system much better because it much more clearly indicates what's expected of the employee, how to make different levels of performance, and so on. However, up and down the line, at various times, we've had some invaluable advice from various people that you don't really need all this fancy nonsense, you just hire a bunch of supervisors who know what they're doing, let them bite the bullet, and give accurate ratings. They can do it. And so, that's generally what I was referring to, that there are people who believe that anybody can do that.

Dr. Kavanagh: Under the new system, will the rating now operate within the system in some way? There will be rewards, punishments, and there wasn't before? I'll buy that because that's the way the civil service appraisal system is in New York. My secretary said "It doesn't matter what you give me as long as it's satisfactory. Who cares?"

Dr. Muckler: I think I find wherever we use ratings as a technique now, increasingly we're asked if we're adding a separate scale which is how much confidence do you have in that rating. And, frankly, the confidence data are much more interesting than the rating data.

Dr. Mullins: What are you doing with the confidence data after you get it?

Dr. Muckler: Mostly going back and saying, "Why did you feel that way?"

Dr. Mullins: You mean just for an interview type thing?

Dr. Muckler: It's really getting back to what's been discussed before--what are the criteria from which these judgments were made? You get a funny pattern. You think certain kinds of things they'd be very confident in and they're not. And conversely. You just want to go talk to people and say, "Why?"

Dr. Kavanagh: When you were talking about your study, Wally, if raters could be making judgments, and then talking about how confident they are about their judgment at that point, that would be a nice addition. In addition, I think that you should be collecting individual difference data in relation to confidence.

Dr. Mullins: John, you're scheduled as a discussant for Wally's paper. Have you already done that?

Dr. Bernardin: I think what I'll do is to combine the two of them; I could do it very easily. One thing that needs clarifying before I start is, what in the world are BARS? We're throwing around that term like we know what we're talking about. BARS are graphic rating scales with behavioral incidents scaled at various points along the graphic scale. We used to call them BES "Behavioral Expectation Scales," and then people chose to completely de-dignify the area and call it BARS.

Dr. Mullins: The acronym is "behaviorally anchored rating scales," BARS. John, I think you're up next.

CHAPTER 3

PERFORMANCE APPRAISAL: SOME NAGGING PROBLEMS AND POSSIBLE SOLUTIONS

H. John Bernardin
Virginia Polytechnic Institute
and State University

After over 5 years of concentrated research in the area of performance appraisal, I feel I'm in a pretty good position to make a reasonably valid statement regarding the effectiveness of appraisal. Regardless of the criterion of effectiveness--discriminability, freedom from contamination, fairness, relevance, convergent, discriminant and external validity, reliability--there is no rating system that is adequate! While at times I felt I was making progress with my little manipulations of behavioral scales with student-teacher samples, field tests of nearly everything I've done have been unpredictable washouts. Now I feel safe in saying that the typical rater is a poor information processor who collects an unrepresentative, unsystematic, and definitely incomplete set of observational data, weights the data according to an invalid stereotype, and then somehow combines it to make a summary rating of performance (a similiar statement has been made about the typical interviewer). The source of the problem is a function of both rater ability and rater motivation. This paper will discuss some possible causes for the lack of rater ability and motivation, critique a popular remedy to the problem (viz. rater training), and suggest some untested approaches to really improving things.

The first section of the paper will question the validity of dimensionalizing work behaviors and the rater's ability to adequately distinguish behaviors by the same ratee. I pick especially on behaviorally anchored rating scales (BARS) but I think the points apply to most rating systems. Rater training has been suggested by many (e.g., DeCotiis and Petit, 1978) as a means to improve rating validity, and the second section will question the most popular methods of rater training and suggest alternatives. The third section will propose an alternative method of assessing performance which calls for separating observation from appraisal, and the paper will conclude with some suggestions for dealing with what I think is the biggest problem in rater motivation--the reluctance to be critical.

Performance Dimensions: Are they real?

There is now fairly strong evidence that respondents will apply a theory of conceptual likenesses in ratings as if it were a theory of behavioral happenings. Items considered to be alike in concept are judged on a rating form to be characteristic of the same person even when the conceptual relationships among the items do not correspond to actual behavior. The "retranslation" phases of BARS development, which is the analogue to item or factor analysis for summated scales, is an exploitation of the theory of conceptual likenesses. Thus, dimensions or traits are generated that are operationalized by critical behaviors thought to be conceptually alike. A number of these critical behaviors (usually 5 to 7 on a 9-point scale) then define the performance continuum and typically the rater is asked to select the item "most typical" or "more representative" of a ratee's performance (see Atkin & Conlon, 1978). Herein lies the problem with the format. For the typical rater with inadequate information on the ratee, the order of items in reference to the ratee will probably remain stable and the rater will be able to make a reliable and invalid rating based on his/her conceptual theory of likenesses. For the rater with adequate information to rate, however, a pre-existing understanding of "what is like what" enters less into the judgment, and the order of items and the similarity of items breaks down leaving the rater with a "scale" of items not necessarily related to one another with reference to particular ratees. Thus, the most typical rater can't make valid ratings because he/she lacks information and the rater with adequate information is probably baffled by the scales. The problem manifests itself again when ratings are made across the dimensions of BARS, on simple graphic scales representing conceptually independent dimensions, or over items on a summated scale. Particularly from raters with inadequate information, the theory of conceptual likenesses will result in dimensions or items with spuriously high correlations. This is of course a new interpretation of halo error. The distinction is that this explanation questions the initial assumption of traits that underlie both measurement and theoretical conceptions of halo. Johnson, among others, questioned even the existence of halo, stating that high correlations that have been found in most studies across traits are probably characteristic of the people being rated rather than a function of errors in rating. He failed to consider the possibility that the correlation could be simply the result of pre-existing conceptual schemes of the raters before they even observe behavior. There is a substantial body of literature that now shows that these conceptual schemes are more highly correlated with ratings of behavior than are the ratings of behavior with actual behavior (see Shweder, 1975). The formation of these conceptual schemes is related to our cognitive tendencies to confirm what is "sensible" and to forget, fail to retrieve, or reinterpret that which is at variance with what is "sensible." This is related to what Johnson-Laird (1972) called "bias toward redundant verification," Garner (1966) called "good form" or "cultural sense" drift, and Tversky and Kahneman (1974) called "representative heuristics" in judgments under uncertainty. The usual explanation for halo is that raters tend to rate individuals according to a general, overall impression of them. Most research on halo has concentrated on scale development, content, and format to eliminate the error. In general, studies concerned with such issues as to whether the position of the "good" end of the scale or whether trait names only versus definitions affected halo have found negative results. No studies have questioned the validity of the traits in the first place despite the fact that studies involving increased acquaintance or observation result in lower intra-trait item correlations along with lower inter-trait correlations.

DeSoto (1961) was probably the only researcher to question the general-impression hypothesis for halo. He argued that people have a predilection for single orderings (i.e., that people dislike multiple orderings). These raters will attempt to achieve consistency by reducing discrepant orderings on ratees across traits to a single ordering. He hypothesized that multiple orderings may be too difficult "to handle" and thus single orderings are sought. Underlying this hypothesis is the assumption that traits are discriminable--thus, the perceived multiple orderings. In fact, he could have used the multiple ordering hypothesis to explain the cognitive formation of traits in the first place.

Symonds (1925) asserted that halo is the strongest on traits that are not easily observed and not clearly defined. He could have just as easily stated that correlations of rated behaviors are highest when behaviors are not easily observed and are not clearly defined. The formation of traits and the rating of traits as, for example, in BARS procedures may be at least partially a function of the unwitting substitution of a theory of conceptual likenesses for a valid description of actual behavior. This may account for the general lack of significant differences between the complicated BARS format and much simpler formats. The behaviors anchoring each BARS are a direct function of the theory of conceptual likenesses and thus only feed the use of the theory in rating actual behavior. A "Barsian" at this point may attempt to rebuke the argument by stating the BARS approach results in behavioral dimensions, not traits and therefore, no theory of conceptual likeness exists. In fact, particularly in the literature on performance appraisal, the distinction between traits and dimensions is unclear (Landy and Farr, for example, have a "behavioral" dimension on their police scales called "attitude"). More importantly, whether we're dealing with traits or dimensions, the theory of conceptual likenesses still pervades the ratings. Newcomb (1929) presented evidence that randomly selected pairs of behaviors were, in reality, as highly associated as pairs of behaviors from any alleged "trait." Schweder (1975) lists numerous other studies dealing with hypothesized behavioral dimensions or patterns. I am not saying behavioral categories cannot be induced from observational records across contexts. I am merely saying that starting out with traits will result in traits that may merely be illusory.

Related to this, the trait approach has been making a comeback of late that is disturbing to me. Two recent studies (DeCotiis, 1977 and Schneier, 1978) found trait-based scales to be psychometrically superior to behaviorally-based formats. The two recent reviews of the performance appraisal area also indirectly endorse or at least sustain the approach. Kavanagh (1971) has of course maintained that if traits can be shown to be construct valid, they should be included in an appraisal system. However, given the procedures that are used to assess construct validity, it is entirely possible that "traits" such as race, sex, height, and weight would end up as appraisal dimensions given the criteria. The argument above notwithstanding, traits are once removed from actual job behavior and twice removed from performance. In an appraisal system, we're seeking valid measures of past performance, not correlates of that performance.

The position I'm getting to is that we shouldn't dimensionalize behavior as with BARS and many other approaches. I am more emphatic on the position if the dimensions are closely akin to personality classifications (e.g., motivation, integrity, coooperation, judgment, attitude, leadership). Such dimensions may be formed by illusory correlations between behaviors with little consideration to contextual interactions. I object less to the behavioral groupings that are done through what is known as "qualitative cluster analysis." However, it would be preferable to induce the groupings based on empirical relationships between observed behaviors of the same ratees across situations rather than through the intuition of psychologists. For example, Dunnette and his associates generated a great number of critical incidents for patrol officer and qualitatively clustered some of them into a grouping or factor known as "using force appropriately." Following retranslation and ratings of incident effectiveness, a scale of the above was developed consisting of eight incidents defining the domain of effectiveness on the alleged factor. Thus, based on its definition, the fact that an officer is more likely to use physical means to arrest a subject is suddenly a correlate of the unjustified use of firearms. My point is that the two behaviors may be totally uncorrelated and the factor called "using force appropriately" may be strictly illusory. This grouping process has a tendency to obfuscate any potential contextual variance. I do recognize, however, that such a grouping has a much greater potential for validity as a grouping than a factor like "attitude" (Landy & Farr, 1975), which was surprisingly developed using essentially the same procedure as Dunnette and his colleagues used.

Rater Training: Are We Merely Training Response Set?

A few years ago, as part of a police validation project, I collected ratings from behaviorally anchored rating scales which surely broke the world's record for skewness. In the hopes of salvaging something from the project, I then embarked on the development of a program to "fix" the ratings. A short time later, all raters (police sergeants) were brought together and grilled on the virtues of bell-shaped curves and variability of ratings across work dimensions and within ratees. The program was preceded by an authoritative endorsement by a high ranking police official. I called this "rater training" and the result was a leptokurtic distribution of ratings replete with what has been called central tendency and median correlation across dimensions of .20. I concluded that rater training had a strong positive effect on rating behavior and that psychometrically superior ratings can be expected from raters who participate in such programs. The results were probably most responsible for my involvement in later studies on the effects of rater training.

Numerous authors have since called for rater or observer training programs to improve ratings of performance. Indeed, several well controlled studies have now shown that rater training can reduce common psychometric errors such as halo effect and leniency, at least as they are measured. I have used several different training programs, in controlled studies and field work, most of which involved the presentations of definitions, graphic illustrations, and examples of common psychometric errors such as leniency, halo effects, and central tendency. Several other studies have used similar strategies.

In his recent review of observer training programs, Spool (1978) concluded that studies assessing training show ". . . accuracy in observation can be improved by training observers to minimize rating errors" (pp. 866-867). Accuracy has been defined as the degree to which ratings are relevant to or correlated with true criterion scores (Dunnette & Borman, 1979). In fact, only two published studies regarding rater training for performance appraisal have used accuracy as a dependent measure. Results indicated in both studies that rating accuracy was not improved as a function of rater training.

None of the other studies concerning performance appraisal and reviewed by Spool directly evaluated rating accuracy or validity. It is only assumed that the more valid ratings are those that have less psychometric error. Typically, the dependent variables used in rater training studies have been leniency error, halo effect, and central tendency. Leniency error is usually operationalized as high mean ratings across dimensions or tests of skewness. Halo error is typically operationalized as either high dimension intercorrelations, low standard deviations across dimensions and within ratees, or significant ratee x dimension interaction effects in an analysis of variance. The extent of central tendency error is usually defined by tests of kurtosis. Less frequently used dependent variables have been interrater agreement, convergent and discriminant validity, and similar-to-me, contrast, and first impression errors. Similar-to-me error is the tendency of the rater to judge more favorably those perceived as most similar to him/her. Contrast effect is the tendency of the rater to evaluate a ratee in comparison to the performance of an employee rated previously. First impression errors occur when a rater evaluates primarily on the basis of initial data.

In a discussion of the various criteria used in the assessment of training programs, Borman (1979) clustered the criteria into three classes. The first class is concerned with what he called "rating behavior" and is operationalized in terms of rating distributions or dimension intercorrelations (e.g., leniency error, central tendency, halo, and range restriction). The second class of criteria is concerned with more crucial attributes of ratings such as convergent and discriminant validity and interrater reliability. The third class of criteria is called accuracy and it encompasses all of the various components of accuracy. Cronbach (1955) has suggested the different accuracy components (elevation, differential elevation, differential accuracy, and stereotype accuracy) will yield uncorrelated accuracy scores. Borman (1979) has argued effectively that the differential accuracy measure is most appropriate for studies on performance appraisal. Of course, field tests of rater training effects usually preclude the assessment of rating accuracy because "true" scores are unavailable.

It is evident from the review by Spool (1978) that most studies on rater training have assessed only one or more measures from Borman's first class of criteria. As stated above, a variety of training methods have proven successful in reducing the first class errors of leniency, halo, and central tendency, as they are statistically defined. The few studies that have investigated variables from the second and third classes of criteria have generally not shown positive effects for training. One exception, Bernardin and Walter (1977), found greater interrater reliability for a group of students who were trained in psychometric error and asked to maintain observational diaries throughout a semester on behaviorally-based scales. Another study (Borman, 1975) found lower interrater reliability in ratings from a group trained on halo effect.

While rater training programs have differed with respect to some key ingredients (e.g., level of participation; practice with the rating scales; feedback to raters), there is a common training core to almost all programs. This core is mainly concerned with changing rater response distributions. For example, in one of the most detailed training programs (Borman, 1979), ratee performances are shown on videotape and trainees rate them. Ratings are then placed on a flip chart and rating distributions are compared and discussed by trainees. A trainer then discusses each error being studied. In the shortest and perhaps simplest training program, Borman (1975) defined halo error and presented a rating distribution indicating the error. The common core to these programs is the presentation of certain rating distributions as an indication of rating error. This same core can be found in almost all other rating training programs. Implied in these programs is that certain rating distributions are desirable while others are not. For example, negatively skewed distributions are considered an indication of halo "error" and raters are encouraged to spread out their ratings for the various dimensions in evaluating a single person. In my field experience with such training programs, I have encountered more than occasional skeptics who do not accept this informal, forced distribution approach. They apparently do not follow the logic that after normal rates of work attrition and some form of personnel selection system, normal distributions of work performance should still be expected and that despite common motivational and/or attitudinal components, ratings across dimensions should be uncorrelated.

In spite of the skeptics, as mentioned above, changing the distributions of ratings through training with this common core has proven successful in both field and laboratory studies. Borman (1979) has stated that getting persons to "spread out their ratings" or to "provide fewer high ratings," is a fairly straightforward procedure but teaching them to rate more reliably or more accurately may well be more difficult. Indeed, the limited evidence available indicates training to enhance accuracy or interrater reliability has not succeeded. Furthermore, the assumption that lower levels of halo and leniency "error" are related to higher levels of accuracy and reliability has not proven to be true. In fact, Borman (1975) found his brief lecture on halo increased variability across dimensions, but the new variability actually decreased rater reliability in identifying ratee strengths and weaknesses.

A tenable hypothesis with regard to rater training is that training on psychometric error as described above merely facilitates the learning of a response set in rating behavior that results in lower mean ratings (i.e., less leniency) and lower scale intercorrelations (i.e., less halo). This new response set may be easily established in the raters given the experimental context of almost all training studies. Research on experimental demand characteristics indicates that subjects in experimental settings respond to even subtle cues in order to respond in a manner they consider to be consistent with an experimenter's expectations. And observer training is anything but subtle regarding the purpose of the "experiment"! A study I just completed strongly supported the hypothesis using vignettes of performance.

In Bernardin (1978), participants in a training program were asked what they had learned from the training. One student wrote "don't give too many high ratings . . . stick close to the middle of the scale for the average. . . don't rate a person high or low on all factors." Many other students responded similarly. And, of course, this training was a temporary "success."

Related to the notion of demand characteristics, Warmke and Billings (in press) assessed the generalizability of effects from three training programs by unobtrusively collecting administrative ratings made by raters about 2 months after training. Higher levels of halo were found in the administrative ratings compared to those collected experimentally and no differences in error rates were found between any trained groups or the control group. While there are other possible explanations for these results, it could be that the response sets fostered by the demand characteristics of the experimental settings diminished or disappeared in the organizational context of the administrative ratings. This is of course in line with the importance of situational context on response set tendencies.

Since most rater training studies do not assess accuracy or validity, it is not known whether the distributions of ratings following the training reflect true score variance or merely response set. Returning to the police validation study I alluded to earlier, while leniency "error" (as defined) was certainly reduced following the training, perhaps had I then concentrated my efforts on the central tendency error, the result may very well have been a return to skewness or perhaps a bimodal distribution! As stated above, this alleged "response set" is of course not occurring in a vacuum. DeCotiis and Petit's (1978) model of the performance appraisal process emphasizes the importance of the organizational context on rating behavior. Bernardin (1978) illustrated the importance of rating context in a simple manipulation. Members of one group of student raters were asked to complete their ratings after completing two short questionnaires. An attempt was made by the experimenter to convey a notion of unimportance in completing the ratings. While passing out the rating scales, the experimenter stated that he "didn't know why the university bothered with such nonsense. The ratings aren't used for anything anyway" (p. 304). Ratings from this group were significantly more lenient than ratings from members of a control group who were completing the ratings for experimental credit.

Prior to training in the police validation study, raters were probably operating on a "response set" that called for lenient, uncritical ratings. There was virtually no pressure from above to rate in any particular manner. Given no pressure to do otherwise, the average rater will tend to be lenient for a number of reasons. Bass (1956) has stated that: 1. The rater may feel that anyone under his jurisdiction who is rated unfavorably will reflect poorly on his own worthiness. 2. He may feel that anyone who could have been rated unfavorably had already been discharged from the organization. 3. He may feel that a derogatory rating will be revealed to the ratee to the detriment of relations between rater and ratee. 4. He may rate leniently in order to win promotions for his men and therefore directly increase his future control of his subordinates by earning a reputation as a superior with "influence upstairs." 5. He may be projecting. 6. He may feel it necessary to always approve of others in order to gain approval for himself. 7. He may be operating on the basis: "Whoever associates with me is meritorious: therefore I am meritorious." 8. He may rate leniently because there exists in the culture a response set to approve rather than disapprove (pp. 359-360).

The "training" program changes things considerably. Essentially, the trainee learns that it is good to be critical. A representative of a higher authority is now telling the raters that rating everyone high will reflect poorly on his/her ability to appraise, suddenly a crucial aspect of supervision. Thus, a new response set replaces an old one. This is particularly true when items 3 and 4 above are not applicable to the rating situation. As the recent data seem to indicate, there is no reason to believe this new response set will result in more reliable, accurate, or valid ratings. I don't believe it was just coincidental that the validation study was a washout, despite the increased variance in ratings following the training and less leniency and halo error.

Thus, we simply don't know if we're training raters to be more accurate or merely "training" response set. My most recent study seems to support the latter. Thus, the significant correlation I reported earlier (Bernardin, 1978) between the internal criterion of the training program (a test of the various types of psychometric error) and the external criterion (the "errors" as defined in the ratings) could be nothing more than an indication that the response set is firmly rooted in the raters.

Suggested Areas of Emphasis

Now that I have eliminated the core of most rater training programs, where does that leave the state of the art in rater training? I still believe that certain types of rater training can inhibit error and increase the validity or accuracy of ratings. A detailed discussion of two approaches will follow. First, let us consider training in the context of a model of the performance appraisal process. Adapting a theory of interpersonal judgment from Taft (1955), DeCotiis and Petit (1978) state that the accuracy of performance appraisals are a function of: a) a rater's motivation to appraise accurately; b) the job-relevance of the standards used by the rater; and c) the rater's ability to evaluate ratee job behavior. Training on rating distributions as discussed above appears to be directed principally at raters' motivation and ability. More specifically, the usual training on errors such as leniency and central tendency are directed at motivation while training on halo, first impression error, and similar-to-me error appears to be directed more at rater ability by making the rater aware that such errors are common. Such awareness should then increase the rater's ability to avoid the errors and perhaps to rate more accurately.

Increasing Observational Skills to Increase Ability-Diary Keeping

To sharpen abilities, Borman (1979) recommends standardizing the observation of behavior and developing a common rater frame of reference for identifying effective and ineffective performance. This two-fisted attack would work on both b) and c) of DeCotiis and Petit's model. It has been shown that behaviorally-based scaling procedures can be used for the development of stereotypes of good and bad workers. The use of a formal diary-keeping system is one way the observation of behavior could be standardized. Bernardin and Walter (1977) trained student raters to maintain critical incidents of instructors' behavior throughout a semester. While the training program also entailed concern for rating distributions and the relative and additive effects of the various parts of the training have not been tested, results did indicate that ratings from the group who maintained observation diaries had significantly less leniency and halo effect and, most importantly, greater interrater agreement than a group of untrained raters. We can indirectly assess the effects of diary-keeping on rating behavior by comparing these results to the results of a later study which used essentially the same training less the diary-keeping. Bernardin (1978) found much weaker effects for this training when instructors were rated with the same behaviorally-anchored rating scales as in the earlier study. Also, 18 of the 20 student raters comprising the diary-keeping group in Bernardin and

Walter (1977) reported in a post-rating questionnaire that the diary-keeping function was "very helpful" in rating instructors. This was signficantly higher than ratings on other aspects of the training program.

It is my contention that a formal system of diary-keeping and observation should be installed after training on critical incident methodology. My field experience with diary-keeping and some research results indicate a formal system may be the only type that works. By formal system, I'm referring to the recording of "x" critical incidents for each ratee over a set period of time. The specific number of incidents depends on the type of job and the observability of the ratee.

A formal system of diary-keeping to be monitored by the rater's supervisor will indicate to the rater that the observation of ratees' behavior is an important job function not to be ignored and that the most important part of the appraisal process takes place all through the appraisal period rather than in the 10 minutes when ratings are actually done. Borman and Dunnette (1975) recommended a closer correspondence between observation and actual ratings. A formal system of diary-keeping and the use of the diaries by the rater to summarize a ratee's performance on a rating scale would seemingly accomplish that recommendation. An alternative approach will be presented below which is probably unworkable but I believe potentially more valid than any other I know of.

Support for a formal system of diary-keeping can also be garnered from differences found in two studies that used diary-keeping. Bernardin and Walter (1977) closely monitored their diary-keepers for frequency and quality of critical incidents generated and the result was diaries qualitatively superior to those that were maintained in a later study where monitoring was far less frequent (Bernardin, 1978). Raters in the earlier study also reported the diaries were a greater help than did raters in the later study. Thus, it appears a rater-student or rater-supervisor will find better things to do unless he/she considers diary-keeping to be an important function. Several studies have shown that the general effectiveness of an appraisal system is a function of the frequency and relevance to performance of the contacts between raters and ratees. Making managers or supervisors aware that their appraisal duties will be assessed as an important job function and monitoring diary-keeping functions should enhance their observational skills.

Recent litigation regarding performance appraisal also indicates the need for something like rater diary-keeping to justify ratings. In Allen v. City of Mobile (1977), the court ruled that police sergeants must justify their ratings of officers with written narratives.

Separating Observation From Appraisal and Appraisal From Appraisee

Since it's so easy to make proposals, let me try one more: why not just do away with a summary rating procedure and somehow convert the diaries to numbers? For example, a supervisor, peer, and/or subordinate enters a set number of real critical incidents per week for each focal ratee. These incidents should be as descriptive and non-evaluative as possible (this will probably require a good deal of training). There should be no mention of traits or dimensions, merely behaviors and contexts. These incidents could be easily entered immediately on a terminal by the observer. Once a fair number of these incidents is entered for each focal person and, preferably, from more than one observer perspective, independent groups of raters, thoroughly familiar with the focal positions, receive randomly ordered lists of these anonymous incidents. Ratings of effectiveness and importance are made for each incident, and descriptive statistics are compiled for each incident, for each focal person, and from each rater. A focal person's rating for each observation period would then be compiled by taking the mean or median effectiveness rating for the group of incidents applicable to him/her, perhaps using importance ratings as multipliers. A paired-comparison approach may be superior here but I won't get into that. While rater bias is somewhat controlled by random assignment of incidents to raters, corrections could also be made as a function of idiosyncratic effectiveness ratings from any one rater's perspective. Judgment analysis could be used to thoroughly study each rater's rating strategy. While there could of course be bias in the recording of the observed incident, the use of more than one observational perspective and the ratee's (it should be observee's) approval of entered incidents for a given observational period should alleviate bias to an extent. Ratees could also be given an opportunity to enter their representative incidents, given a consensus agreement with the supervisors over language.

It can of course be argued that there is evaluation and interpretation in any behavioral observation. However, while I agree with this statement, it must at least be recognized that something akin to immediate scoring observation or description is much more detailed, systematic, and reliable than ratings. If, for example, immediate scoring observation would adopt a real critical incident methodology, I would wager that the results would be far less evaluative and biased than any rating format. I emphasize a real critical incident methodology because this term has been butchered under the guise of BARS. While I could cite numerous examples, consider this "critical incident" on the Landy and Farr (1975) police BARS for the dimensions "work attitude": "only goes through the motions of the job" or another, "always does his share of the work." I have a strong suspicion that this is not what Flanagan had in mind. What appears to be lost in most alleged critical incidents used on BARS is the all important description of the context for the behavior. Without a detailed description of the context, we don't even know if there were behavioral alternatives for the ratee (i.e., whether he/she could have done something else). This may be related to what has been referred to as presumptive bias in favor of main effects; that is, we ignore context differences in which behaviors were observed and conclude that different behaviors are simply a function of some stable behavioral pattern, characteristic, or trait difference. Schweder (1978) has stated that a world of complex multiple necessary causes, and person by context by response mode interaction effects is not one that judges are inclined to consider when they draw inferences about individual differences.

Thus, I believe if we separate description from ratings and ratings of specific behavior from ratings of behaviors by certain people, we will end up with a more valid numbering system for individual differences, relatively free of the bias affecting any appraisal system as we know it today. This approach is obviously far more cumbersome than a basic 6-month review procedure. Supervisors will undoubtedly hate the system at the outset due to the added time for more detailed observation, then entering incidents, and finally rating anonymous incidents. They may also object to the fact that the control of numbers for their subordinates is now out of their hands (much like the reaction to forced-choice methodology). Additionally, the whole thing may smack of Theory X. The system may also be unworkable if computer terminals are not easily accessible. However, in terms of valid employment decisions, I believe the pay-off could be great. The compilation of critical incidents could also result in the development of reality-based behavioral groupings that take context into consideration.

Validate Raters Not Instruments

Returning to reality for a moment, the system could also be used on a one-shot basis to "validate" raters. Raters would be asked to follow the same procedures as above and, in addition, do standard ratings on their ratees. Judgment analysis could be run on their rating policy and correlations could be run between "anonymous" incident ratings within ratees and ratee ratings by the supervisor. This could be construed as a form of construct validation for a particular rater and this is a critical point, and time for another digression. Several writers have discussed the validity of, for example, BARS relative to other formats and in an absolute sense. Schwab, Heneman, and DeCotiis (1975), among others, have stated BARS have content validity because dimensions and incidents are generated that define the work domain. While the procedure for developing BARS certainly can be considered content-oriented, it is erroneous to refer to any rating instrument or format per se as content valid. Research using Landy and Farr's (1975) police scales illustrates the point most dramatically. The authors found big differences in measures of convergent and discriminant validity in ratings from their BARS across the various police agencies sampled. Thus, not even particular scales (let alone formats) possess certain levels of reliability or validity but rather the resultant ratings from specific raters do or do not possess these qualities under certain circumstances. Guion (1977) has stated that we cannot assume that we have a valid number system for any instrument merely because we have a representative sample of content. This statement must be applied emphatically to performance ratings with their great potential for contaminating sources of error. Borrowing from Guion (1978b), carefulness in scale construction or content sampling should not be mistaken for validity. The fact that a rater is using BARS as opposed to simpler formats does not insure that ratings will be any more valid or any less contaminated than ratings from any other format. I'm now convinced that we can help the conscientious rater to some extent with format but the "screw off" will do so on any format he/she pleases. Thus, we must validate raters, not instruments.

I believe the evidence is now strong that most judges are not cognitively prepared to adequately summarize and abstract from a great many observations. This is of course the essence of performance appraisal. Raters are asked to summarize sometimes a year's worth of observation into a definitive rating. It is also fairly clear that raters cannot document very well the basis of their rating. My early research with BARS found that psychometrically superior ratings would result when raters are asked to record and scale a minimum of three critical incidents per dimension per ratee. My subsequent field tests of this approach reveal that raters, even those with frequent opportunities to observe ratee behavior, cannot document what they have observed very well at all. Students, for example, when asked to generate as many critical incidents as they could after observing an instructor for 3 hours a week for 15 weeks, came up with an average of about four statements per instructor, the mode of which lacked sufficient detail to be called a critical incident. This can be improved somewhat with rater training and surveillance over the observational period, as discussed above, but even with these conditions, the number of behavioral incidents ratees can retrieve at the same time they're doing summary ratings is very low. And there is strong evidence in the information processing literature that indicates what is retrieved is not representative of ratees' behaviors anyway. Thus, the inference that a rater is drawing when doing summary ratings for an observational period is probably drawn from very limited and unrepresentative recall of ratee behavior. A student, for example, once rated an instructor very low on "student-teacher rapport" and wrote "he didn't recognize me in the hall one day." A police sergeant rated one of his officers very high on "judgment" and wrote only that "he covered an exit ramp to a highway leading out of town once when there was a robbery in progress." While this description could in fact be representative of the officer's good "judgment," why couldn't the sergeant cite other examples? It must be correct to assume that the rating on judgment is based almost exclusively on the recall of this one event and the potentially invalid inference made from it. You might argue that the incident is recalled because it fits with a prior attribution regarding the officer's "judgment." However, could this not be just another case of "bias toward redundant verification" or perhaps selective perception?

Thus, in terms of valid individual differences between workers, I say we must change appraisal systems to predominantly observation systems. Due to the bias inherent in any rater, we must also separate the person from the rated behavior. A procedure to accomplish these two things is outlined above. Anything approaching it will probably be better (i.e., more valid) than the more traditional appraisal approaches.

Training Raters to be Critical: A Social Learning Conceptualization

Returning to the typical appraisal system and its typical problems, perhaps it would be propitious to conceptualize the tendency of raters to be leni^nt as essentially defense behavior. McGregor (1957) and many others have discussed the reluctance of evaluators to "play God" as it were. While training directed at improving observational effectiveness may be necessary for more accurate ratings, such training is probably not sufficient. If we consider the tendency to be lenient as a defensive behavior (i.e., avoiding the reactions of ratees to harsh ratings), what we need are psychological methods that create and strengthen the expectations of personal efficacy. In the classic work of Bandura (1977b), an efficacy expectation is the conviction that one can successfully execute a behavior in order to produce a certain outcome. Efficacy expectations have been distinguished from outcome expectancy which is conceived as the estimate that a given behavior will lead to a given outcome. This distinction is critical in considering the cognitive processes of a typical lenient rater. The rater could very well believe that a firm, albeit harsh rating will get a subordinate going or will be the basis for a critical administrative decision (the latter outcome can be easily changed by the organization). However, the rater could seriously question his/her capability of coping with the resultant situation (e.g., the ratee's rage). Bandura has stated that the strength of convictions in one's own effectiveness determines whether coping behavior will be attempted in the first place. People fear and avoid threatening situations they believe exceed their coping abilities, whereas they behave assuredly when they judge themselves capable of managing situations that otherwise intimidate them

So let us consider next what psychological methods could be used to establish and strengthen self-efficacy in performance appraisal. Bandura has presented four main sources of information which facilitate personal efficacy. Performance accomplishment is considered the most influential source because it is based on experiences of personal mastery. It is probably through failures in making fair but critical performance appraisals early in a manager-supervisor's career that account to a large extent for low levels of personal efficacy. Expectations are also derived from vicarious experience whereby a person observes someone coping with problems and succeeding. Spool (1978) has recommended a modeling approach to rater training in which trainees observe model persons behaving in appropriate ways. However, a large body of research indicates modeling is a less dependable source of information than is personal accomplishment. Thus, efficacy expectations facilitated by vicarious experience will be weaker and more subject to change. Latham, Wexley, and Pursell (1975) used a modeling approach where trainees observed fictitious, videotaped managers making observational errors. The psychological distance between this contrived situation and the contextual realities of an ongoing performance appraisal system render the practical effectiveness of this approach tenuous.

Verbal persuasion is the third source of information for expectations of personal efficacy. This method consists of essentially "coaching" persons into believing they can cope. Lacking a real experimental base, this approach is also weak compared to actual accomplishments. Finally, emotional arousal can change efficacy expectation in intimidating situations. Anxiety over the repercussions of negative feedback could debilitate a rater's ability or motivation to give accurate ratings. Bandura (1977a) and Sarason (1975) have proposed several methods to eliminate defensive behavior by diminishing emotional arousal.

As far as selecting the best psychological method for changing lenient behavior, research seems to strongly support the use of performance treatments designed to master experiences. Translating this into a performance appraisal system, perhaps something like the following could be tried. First, an instrument assessing perceived self-efficacy could be developed where efficacy expectations could be measured. A short list of performances could be presented dealing with giving subordinates negative feedback when justified (i.e., telling a subordinate with whom you socialize that the subordinate's tardiness record is getting out of hand). Respondents could indicate the strength of their expectations on a probability scale. Specificity and generality could be worked into the scale by having respondents focus on both specific subordinates with whom they do or do not socialize and subordinates in general. The scale could be validated using an external criterion involving an opportunity to provide justifiable negative feedback. Efforts must also be made to determine if we're merely dealing with a new measure of assertiveness.

After the refinement of the self-efficacy expectation scale, scores on the scales could be used to determine what (if any) training is necessary for each respondent. Real-life behavior relating to feedback could also be scaled for difficulty and persons could be told to perform certain behaviors based on their expectation score. For example, a person who scores very low on expectations could be given relatively easy "homework" assignments regarding negative feedback such as interviewing a subordinate whose work is exemplary with the exception of one small rather insignificant area. These specific "homework" assignments should follow standardized training on how to conduct a performance appraisal interview, the details of which are beyond the scope of this paper (see Lefton, Buzzota, Sherberg, & Karraker, 1977, for excellent discussions of alternative methods). Suffice it to say that any of the recommended appraisal interview styles or concomitant rating instruments still require some coping behavior on the part of the rater for which he/she may not be prepared. Little attention in fact is given to rater defensive behavior. These "how to" books and articles are also nothing more than verbal persuasion even if they are directed at self-efficacy expectations. The results of this approach have already been found to be wanting. The incorporation of sources of information through vicarious experience, verbal persuasion, and emotional arousal as precursors to actual "homework" could enhance the probability and persistence of effort. The use of relaxation techniques, for example, immediately prior to the actual performance would probably increase the chances of personal mastery at each level of performance.

The entire training program would be set up in a systematic desensitization format whereby the potentially intimidating performances are broken up into more easily mastered steps of increasing difficulty culminating in the necessary encounter with the greatest difficulty (e.g., telling a subordinate friend his/her work is incompetent); there are times when such harsh judgments are necessary despite what is said about the goal-setting approach.

BIBLIOGRAPHY

Allen v. City of Mobile (331 .F. Supp. 1134, 1977).

Atkin, R.S., & Conlon, D.J. Behaviorally anchored rating scales: Some theoretical issues. *Academy of Management Review*, 1978, *3*, 119-128.

Bandura, A. *Social learning theory*. Englewood Cliffs, NJ: Prentice-Hall, 1977. (a)

Bandura, A., Adams, N.E., & Beyer, J. Cognitive processes mediating behavioral change. *Journal of Personality and Social Psychology*, 1977, *35*, 125-139. (b)

Bandura, A., Jeffrey, R.W., & Gajdos, E. Generalizing change through participant modeling with self-directed mastery. *Behaviour Research and Therapy*, 1975, *13*, 141-152.

Bass, B.M. Reducing leniency in merit ratings. *Personnel Psychology*, 1956, *9*, 359-369.

Beatty, R.W. A comparison of the operationalization of behavior-based versus effectiveness-based performance appraisals. Paper presented at the annual meeting of the American Psychological Association, 1977.

Bernardin, H.J. Effects of rater training on leniency and halo errors in student ratings of instructors. *Journal of Applied Psychology*, 1978, *63*, 301-308.

Bernardin, H.J. The impact of role perception on performance appraisal. Paper presented at the annual meeting of the American Psychological Association, 1977.

Bernardin, H.J., & Boetcher, R. The effects of rater training and cognitive complexity on psychometric error in ratings. Paper presented at the annual meeting of the American Psychological Association, 1978.

Bernardin, H.J., & Walter C.S. Effects of rater training and diary-keeping on psychometric error in ratings. *Journal of Applied Psychology*, 1977, *62*, 64-69.

Borman, W.C. Format and training effect on rating accuracy and rater errors. *Journal of Applied Psychology*, 1979, *64*, 410-421.

Borman, W.C. Effects of instructions to avoid halo error on reliability and validity of performance evaluation ratings. *Journal of Applied Psychology*, 1975, *60*, 556-560.

Borman, W.C., & Dunnette, M.D. Behavior-based versus trait-oriented performance ratings: An empirical study. *Journal of Applied Psychology*, 1975, *60*, 561-565.

Brown, E.M. Influence of training, method, and relationship on the halo effect. *Journal of Applied Psychology*, 1968, *52*, 195-199.

Cronbach, L.J. Processes affecting scores on understanding of others and assuming "similarity." *Psychological Bulletin*, 1955, *52*, 177-193.

Cummings, L.L., & Schwab, D.P. *Performance in Organizations.* Glenview, IL: Scott, Foresman & Co., 1973.

DeCotiis, T.A. An analysis of the external validity and applied relevance of three rating formats. *Organizational Behavior and Human Performance*, 1977, *19*, 247-266.

DeCotiis, T., & Petit, A. The performance appraisal process: A model and some testable propositions. *Academy of Management Review*, July 1978, 635-646.

DeSoto, C.B. The predilection for single orderings. *Journal of Abnormal and Social Psychology*, 1961, *62*, 16-23.

Downey, R.G., & Saal, F.E. Evaluating human judgment techniques. Paper presented at the annual meeting of the American Psychological Association, 1978.

Dunnette, M.D., & Borman, W.C. Personnel selection and classification systems. *Annual Review in Psychology*, 1979.

Freeberg, N.E. Relevance of rater-ratee acquaintance in the validity and reliability of ratings. *Journal of Applied Psychology*, 1969, *53*, 518-524.

Garner, W.R. To perceive is to know. *American Psychologist*, 1966, *21*, 11-19.

Guion, R.M. Scoring of content domain samples: The problem of fairness. *Journal of Applied Psychology*, 1978, *63*, 499-506. (a)

Guion, R.M. "Content Validity" in moderation. *Personnel Psychology*, 1978, *31*, 205-213. (b)

Guion, R.M. Content validity: The source of my discontent. *Applied Psychological Measurement*, 1977, *1*, 1-10.

Hollander, E.P. Validity of peer nominations in predicting a distant performance criterion. *Journal of Applied Psychology*, 1965, *49*, 434-438.

Jackson, D.N., & Messick, S. Response styles and the assessment of psychopathology. In S. Messick and J. Ross (Eds.). Measurement in Personality and Cognition. New York: Wiley, 1962.

Johnson-Laird, P.N., Legrenzi, P., & Legrinzi, M.A. Reasoning and a sense of reality. British Journal of Psychology, 1972, 63, 395-400.

Kavanagh, M.J. The content issue in performance appraisal: A review. Personnel Psychology, 1971, 24, 653-668.

Kay, E., Meyer, H.H., & French, J.R.P. Effects of threat in a performance appraisal interview. Journal of Applied Psychology, 1965, 49, 311-317.

Landy, F., & Farr, J.L. Performance rating. Unpublished manuscript. Penn State University, 1978.

Landy, F.J., & Farr, L.J. Police performance appraisal. Law Enforcement Assistance Administrative, 1975.

Landy, F.J., Farr, J.L., Saal, F.E., & Freytag, W.R. Behaviorally anchored scales for rating the performance of police officers. Journal of Applied Psychology, 1976, 61, 750-758.

Latham, G.P., Wexley, K.N., & Pursell, E.D. Training managers to minimize rating errors in the observation of behavior. Journal of Applied Psychology, 1975, 60, 550-555.

Lefton, R.E., Buzzota, V.R., Sherberg, M., & Karraker, D.L. Effective motivation through performance appraisal. New York: John Wiley, 1977.

Levine, J., & Butler, J. Lecture versus group discussion in changing behavior. Journal of Applied Psychology, 1952, 36, 29-33.

Maier, N.R.F. The Appraisal Interview: Three Basic Approaches. LaJolla, CA: University Associates, 1976.

McCall, M., & DeVries, D., Appraisal in context: Clashing with organizational realities. Paper presented at the annual meeting of the American Psychological Association, 1976.

McGregor, D. An uneasy look at performance appraisal. Harvard Business Review, 1957, 35, 89-94.

Meyer, H.H., Kay, E., & French, J.R.P. Split role in performance appraisal. Harvard Business Review, 1965, 43, 123-129.

Newcomb, T.M. The consisting of certain extrovert-introvert behavior patterns in 51 problem boys. Teachers College, Columbia University, Contributions to Education, 1929, No. 382.

Rand, T.M., & Wexley, K.N. A demonstration of the Byrne similarity hypothesis in simulated employment interviews. *Psychological Reports*, 1975, *36*, 535-544.

Rosenthal, R. *Experimenter effects in behavioral research*. New York: Appleton-Century-Crofts, 1966.

Sarason, I.G. Anxiety and self-preoccupation. In I.G. Sarason & C.D. Spielberger (Eds.), *Stress and Anxiety*. Washington, DC, Hemisphere, 1975.

Schneier, C.E. Measuring human performance in organizations: an empirical comparison of the psychometric properties and operational utility of two types of criteria content. *Proceedings of the Academy of Management*, 1978.

Schwab, D.P., Heneman, H., & DeCotiis, T.A. Behaviorally anchored rating scales: A review of the literature. *Personnel Psychology*, 1975, *28*, 549-562.

Schweder, R.A. Fact and artifact in personality assessment: The influence of conceptual schemata on individual difference judgments. *APA*, 1978.

Schweder, R.A. How relevant is an individual difference theory of personality? *Journal of Personality*, 1975, *43*, 455-484.

Spool, M.D. Training programs for observers of behavior: A review. *Personnel Psychology*, 1978, *31*, 853-888.

Symonds, P.M. Notes on rating. *Journal of Applied Psychology*, 1925, *9*, 188-195.

Taft, R. The ability to judge people. *Psychological Bulletin*, 1955, *52*, 1-23.

Taguiri, R. Person perception. In G. Lindzey & E. Aronson (Eds.). *The Handbook of Social Psychology*, Vol. III. Reading, MA: Addison-Wesley Publishing Company, 1969.

Tversky, A., & Kahneman, D. Judgment under uncertainty: Heuristics and biases. *Science*, 1974, *185*, 1124-1131.

Warmke, D., & Billings, R. A comparison of training metho's for altering the psychometric properties of experimental and administrative performance ratings. *Journal of Applied Psychology*, in press.

Weber, S., & Cook, T. Subject effects in laboratory research: An examination of subject roles, demand characteristics, and valid inference. *Psychological Bulletin*, 1972, *77*, 273-295.

Wolpe, J. *The practice of behavior therapy*. New York: Pergamon, 1974.

Dr. Bernardin: I already mentioned a couple things that reply to Wally's paper and I'm going to try to throw in my paper as well here to catch up.

Assessing validity with no accuracy scores, this is where the rubber meets the road problem again. Obviously a construct validation approach is the one that should be taken given that the psychometric approach is not so good, meaning leniency and halo.

I think there is an excellent paper by Kane and Lawler (1979, in Staw's Research in Organizational Behavior, JAI Press, pages 425-478) very recently, which makes some suggestions on how to confirm convergent and discriminant validity that isn't very well taken care of up until that point. The Kane and Lawler article gets into a better way to look at convergent and discriminant validity so that you know how much variance is accounted for by those indices. I think that is probably the best approach right now that we have for looking at something like validity when we have no true scores or no accuracy measures, so I recommend that extension of Kavanagh's work.

Yankalovich did a poll years ago, pre-Watergate, on Nixon when he was very popular mostly because of McGovern's blunders, and they asked the people what they thought of Nixon's tax reform bill and 89% strongly agreed with the tax reform bill. Nixon never did introduce the tax reform bill. Yankalovich only threw the question in just to get an indication of how much halo was running through their feelings of Nixon. Kane and Lawler also talk about a way of looking at that kind of bias. It's a significant rater by dimension interaction and, as implied by Wally's paper and a big review article by Landy and Farr, there may be a difference in terms of implicit theories of how dimensions or traits or factors of performance relate to one another. This gets very close to Pat Smith's definition of halo, as ratings of one characteristic spilling over to affect ratings on other characteristics. There have been a lot of attempts to do away with halo effect and most of them have been failures although most of them have been real dumb little manipulations like where to put the scale points, and adjective to use, and the like.

Some recent work done by Anthony Dalessio (1979, SEPA Presentation) assumes that there are illusory correlations, meaning people have a belief system about how dimensions fit together. It starts with that assumption and exploits it in the rating process by constructing a matrix for ratings. Let's say you have three dimensions, you would have a three by three matrix and you'd ask people to rate a ratee on the interaction of two dimensions, let's say ability to motivate and leadership, if these are two of the dimensions. Rate on the interaction of those until you rate all nine cells of that three by three matrix. And then you'd go down the column and compute an average for the various dimensions.

Dalessio found that that rating procedure did somewhat eradicate halo, not as much as he had hoped, and I think the explanation for that is basically the explanation for halo which is Norman Anderson's, again back to the social research. Anderson's discussion of the general impression theory of person perception, where we make an overall judgment right at the start and that overall judgment affects our ratings or our perceptions of dimensions that are more specific. It may also speak to the discussion we had earlier regarding global and specific dimensions. But if a person makes an overall judgment and later he gets into rating a specific dimension, then there's a regression to that overall judgment so that the means will all be glumped together toward that overall judgment, and I think it's a good explanation for halo.

Also, it might be related to Wally's reference to Vernon's work regarding analyzers and synthesizers. Maybe it's the analyzers which are doing that. They form judgments right after observing behavior whereas the synthesizers aren't doing that. It might be a good study to look at who's forming those initial judgments--and that kind of content analysis that Banks is doing is a good way of getting at that sort of thing. Can we predict from Vernon's typology? Selecting raters, I think, is a good idea. Unfortunately Wally's work is based on tapes and the generalizability is yet to be tested although your research did support pretty much this stuff in the social research, same kind of dimensions, same kind of relationships.

My only problem with that is your studies are relating the ratings made by these people with these individual difference kinds of variables and the ratings made by experts. Are we only perhaps just dealing with the correlation between expert characteristics and rater characteristics? I know I was an expert, Mickey was an expert in a later study, and the study you're referring to was not. Was it graduate students?

Dr. Borman: We've had two different sets of expert raters now.

Dr. Bernardin: Isn't that a possibility? That we're just dealing with the correlation between characteristics of the alleged expert raters who didn't do too well in terms of interrater agreement and the characteristics of the raters as you've measured them? And maybe again accuracy is somewhere off in the wind somewhere. That's one problem that I see with that study in terms of its generalizability.

There's nothing wrong with intelligence, though, as a common correlate. Wally mentioned administrative set. I think that is far and away the biggy, especially in terms of generalizability. There's a study coming out in JAP by Warmke and Billings that goes to all sorts of trouble training raters and getting halo down to a decent level and no leniency effect in these little games that they're playing, rating people anonymously, and no administrative ramifications. And then Warmke and Billings went into the real world, 6 months later, followed these raters and collected their ratings, and there was a total breakdown of their results.

Again, back to the real world, back to the administrative set. I mentioned in my paper the effects that you can have on administrative set may be fairly drastic. I'm not sure what the effect is on validity, but you can affect behavior rather quickly. I did an Ohio validity study and I think I mentioned they broke the world skewness record, they were just abominable ratings on BARS, and so we had all of the raters, who were police sergeants, in and we grilled them on rating distributions of various kinds, and don't commit halo, and don't do leniency. And we also had the Chief of Police give a nice little talk which I think helped things; you know, "Please be cruel" essentially. And the ratings were extremely leptokurtic into the middle range, just a complete change. And this was like 2 months after we'd collected these BAR ratings on the same people. So you can have a strong impact on the administrative set with some type of intervention, as Mickey mentioned.

I haven't said anything about validity, however, in my paper. I'm not sure we were getting any more valid ratings with these scare tactics and don't do this, and don't do that. We did get something closer to what we wanted but I should say in the end that the validity study was a total washout despite fairly decent variability in the criterion; so that may speak to the issue of validity, meaning we have none in terms of those ratings.

Objective data . . . back to the points we raised earlier with Cecil's paper, I'm on Cecil's side in terms of objective data. I think you should use it if it's available, despite deficiencies. I'm sure you're not going to be able to cover the ultimate criteria very well, but if you have it, you should use it, and I don't think the literature is very good on correcting for opportunity bias. If you go back to the limited number of studies that relate objective data to subjective data, you'll find very few corrections for opportunity bias of any kind or contaminations of any kind. Stan Seashore's work (Ronan & Prien's Book, 1971) is the one that's the most often cited; in subsequent work very little attempt is made to correct for opportunity bias. The Kansas City Police Department is one department that has corrected for alleged opportunity bias, meaning patrol assignment, whether you're in the ghetto, whether you're in the suburbs, and they've found that correcting for opportunity bias has increased their validity in terms of predictability of selection instruments using that. So I think that there's a possibility that that may be one of the shortcomings of objective data.

Dr. Cascio: When you say opportunity bias, are you talking about opportunity to observe in class or are you talking about opportunity to demonstrate performance?

Dr. Bernardin: The latter. In my police example, number of arrests leading to conviction, that's a fairly decent criterion. It makes sense.

Dr. Cascio: Oh, very clearly. Several years ago when PDI was doing some work in Miami, when you were developing your behaviorally anchored rating scales, one of the things that came out of that was that female police officers were always rated very poorly. Why is that? And we began to interview people and found out a reason. At that time they had a rule that a male always had to ride with a female. They couldn't put two females in the same patrol car. And when anything really heavy went down, to use the street language, the male always took the lead, and the female never had a chance to demonstrate whether she could or she couldn't perform in that situation. So she really never had an opportunity to perform and consequently raters were being conservative and rated them all low.

Dr. Bernardin: Yes, that's a good example I think. In our police experience, opportunity bias is a real problem, but I think you can get at it, at least to an extent, and it still may be better after you correct for opportunity bias as much as you can; it still may be better than ratings. At least that's been my experience with ratings in the field with police. So that's my preference toward objective data, if possible.

Another applied problem. With our accuracy measure and also with our convergent/discriminant validity presentation, one of the nitty gritties of it all that I think Dr. Muckler is going to get into regarding the Civil Service Reform Act is that when we're dealing with due process, when we're dealing with Title VII, I don't think that data is going to be worth anything. It wasn't in Hill vs. one of the electric companies because you have an individual, a ratee, who is bringing suit, or you have a class. The Black Patrolman's Union in Toledo, Ohio, brought suit against the Toledo Police Department when I was doing work there. A class, maybe a Title VII case. And they're going to find adverse impact fairly simply in a lot of situations, and the burden is going to fall on the employer to come up with a discussion of convergent and discriminant validity which completely dilutes the rater/ratee interaction.

It may not stand up at all in a court of law. You have to validate raters--and this goes back to a point I made earlier. There could be a real bozo rater in there who is responsible for a ratee's lower ratings who happens to be a Black or female wh brought suit.

Convergent-discriminant validity isn't going to do anything for such a case. So, again, validate raters along with the rating procedures and I think you'll be a lot safer.

Some of the things I raised earlier speak to that. Another case--I'm getting a little bit too practical here, but Allen versus the City of Mobile is an important performance appraisal case, just supported in appeal. It required written narratives or justifications of numerical ratings, a point I raised earlier, and also five raters per rating period, two of whom were selected by the ratee. I think we should speak to that issue as well, in terms of setting up a performance appraisal system. That case is going to be quoted often.

As to leniency, I mentioned the deal of getting objective data on some variable and also asking the rater to rate on that variable and correlating the two together. It's a way of investigating whether you're dealing with legitimate true score skewness or whether you're dealing with error. Mickey implied that that's a problem in terms of making generalizations to other dimensions, but it's a start.

Okay, I disagree with one thing Wally said, and also Landy and Farr, who have called for a moratorium on format research. And I disagree for one major reason. The comparisons that have been made between formats (for example, BARS versus traditional graphic scale, BARS versus mixed standards, forced-choice, whatever) have completely confounded formats within the BARS procedure, always in the way they're developed and in the final product and also in the actual rating procedure.

So just a short lesson in BARS development will illustrate my point. Smith and Kendall--I'm not saying they developed the best system or anything--recommended a system where we have three characteristics. At the top is a dimension that's defined in a general sort of way. There are also dimension clarification statements which are just illustrations, very general in nature, of the dimension in question that serve to generally anchor three points on the scale and then critical incidents anchor the scale at various points.

Now there may be a problem with the critical incidents. Wally addresses one study--it may be better to go to more generic explanations for generalizability. But this is the basic format that Smith and Kendall used. They recommended--and this is almost completely ignored in the research except for stuff that I do and stuff that Zedeck (Sheldon Zedeck, Dept of Psych. U/Calif/Berkeley) does--they recommended actually writing critical incidents on the scales, a minimum of two--like the original work with the National League of Nurses--and then using the mean of the newly scaled critical incidents as the rating. It was more of an observation technique. Observation inference, the inference being the relationship between what you have observed, the critical incidents you're generating, and these dimensions.

Dr. Mullins: John, you've got critical incidents on the right of the vertical line. What's on the left?

Dr. Bernardin: These are dimension clarification statements. For example, if you're dealing with leadership skill, this would be a generic explanation of a high level of leadership skill rated high.

Dr. Mullins: Essentially the same thing as a description of the behavior of high leadership and the behavior of low leadership?

Dr. Bernardin: Cxactly.

Dr. Muckler: On top would be "got his men to attack machine guns with their bare hands." On the bottom would be "ran away when first shot was fired."

Dr. Bernardin: That might be a critical incident, which is another amusing thing. If you go through the BARS procedure, you can get incidents that will survive all the steps of BARS that are absolutely abominable.

One example we got, "this officer while on duty went out of his district, went into a bar, got drunk, and had his gun stolen." Now that will survive all the steps of a BARS procedure and just because of ceiling effects, it will not be down here at the very bottom. Some bozo will put a 2 or a 3 on rating effectiveness which completely kills this whole section of the rating scale. So you have to make sure there is some probability that such behaviors will be elicited by those people on the work force, not people who have been fired. I suspect that person was dismissed after such behavior.

Anyway, this is the basic format that I tested in a study in 1976 and using that scoring procedure where they generated actual critical incidents, took the mean of the newly scaled incidents and measured halo effect, leniency error, and interrater agreement, which I think is a little better even though there are some potential, systematic biases. And I compared those to the more basic types of ratings which is the most popular type, "just check a point along the scale and that's your rating," versus some other procedures, and I found that scaling of critical incidents was far and away better than any other rating procedures and that this set-up with the dimension clarification statement and the incidents was also better than a system that didn't use dimension clarification statements.

What's happened is that a good many studies claiming to use BARS have a set-up like this. They may have an 8-point scale with eight critical incidents, one right after the other. This scale is probably going to be a disaster if you consider the overlaps and the distributions of these incidents. When you're thinking of a particular person and putting that person in the place of those situations, they may overlap completely and there'll be a total breakdown of this scale in terms of going from low effectiveness to high effectiveness. And there are no dimension clarification statements.

This is a procedure that's discussed by Atkin and Conlon in the 1979 Academy of Management Review paper, and they say, "this doesn't make any sense and you shouldn't do it this way and BARS isn't worth a hill of beans, etc., etc." Well, I agree with that in terms of these because these assume a Guttman kind of scaling process, and Smith and Kendall's is an interval scaling process. In this procedure here, you must have a high index of reproducibility with this scale, and very often when people develop this kind of scale, they don't even do indices of reproducibility or scaleogram analysis on this. They just develop these incidents. What happens is it's a disaster, and in comparing that disaster to graphic scales or summated scales, the latter come out better or it's a tie, and they conclude that BARS are not any better.

I maintain that the reason is because what's been used under the guise of BARS are some perhaps faulty methodologies and that if you use this procedure, they may come out better. Now after I've said that, I have to say, my field experience, which is completely different from my lab experience, doesn't support that as strongly as I'd like. Even this system, and I think it just speaks to administrative set again, even that system, as great as I maintained it was in this '76 article, is just a disaster if you get into the wrong kind of situation. It's exactly what Wally was talking about, "Is there a desire to rate accurately?"

What I maintained in my paper in the latter section discussing Albert Bandura is that I think maybe in the majority of cases there is not a desire to rate accurately. There is perhaps a need to avoid a confrontation with subordinates so you tend to be more lenient. And also there's just other sets operating that are like Bass's (1956, Personnel Psychology) discussion of leniency effect and some of the things that go through a rater's mind; the context, what it means in terms of the rater's position in the organizational hierarchy. If he or she rates a person high, is that person going to be promoted? Will that rater then have more influence in the organization at large against some crucial issues which might fall under the rubric of administrative set and also personal set? I think that is more important perhaps than format.

Dr. Mullins: John, do you know of any instances where anyone has constructed a scale using the Smith and Kendall technique very purely and scientifically and compared that with ratings constructed otherwise in a situation where outside validity is available? Do you know of any advantage that this has shown in terms of being valid?

Dr. Bernardin: Not with external validity, no. Latham and Ronan did some stuff with loggers where they broke the scale down into summated ratings; they took these things and just made them into summated scales, not dimensionalized, and they found a fairly strong relationship to logger performance, right? But that's the only study that I can think of that uses even behavioral scales that were developed using at least a variant of Smith and Kendall's methodology. I haven't. I wish I had but the objective data in my police studies is always so screwed up that I can't use it.

Dr. Kavanagh: Have you thought of using the approach of constructing forced-choice ratings and comparing BARS against forced-choice ratings?

Dr. Bernardin: Yeah, I did. I have just developed forced-choice ratings, as a matter of fact, and I followed the steps Wherry suggested and other people who researched forced-choice and they're still fakable. I mean you get high levels of racial and sex bias in the forced-choice scales and there were no differences between any of the formats I tested. I have used summated scales in comparison. The last thing I wanted to say is I did a study in '76 comparing BARS to summated scales. The summated scales were much better on all the variables. I then went through a more painstaking developmental criterion for the BARS approach and found that they were significantly improved compared to similar samples of people in a later study and no differences between summated scale and BARS. However, another criticism of Wally's paper, that was one study among many studies that used raters who were rating in the same setting the same ratees across two formats. Meaning they finished the BARS and went right on to the summated scales and I'm sure there would be a desire to be consistent in that kind of a methodology, and that may account for no signficant differences. If you could somehow randomly select ratees and raters and compare formats, you may get big differences. Do you know of a study that used that approach?

Dr. Borman: The counterbalanced kind of study or an approach where the rater uses one and only one scale? Ours have been counterbalanced, essentially. That's one way to get around it. Half the people have the kind of advantage you're talking about; the others have the opposite kind of order. No, I don't know of any other work like that.

Dr. Bernardin: That may account for the Schwab, Heneman, & DeCotiis (Personnel Psychology, 1975) conclusion that there are no format differences because every study they cited at that time was that kind of a set-up.

Dr. Ree: Before you leave the topic of format, you made the statement that this beanpole on the right with the Guttman scale procedure was inferior in theory and in laboratory practice to the BAR procedure on the left, and I'm at a loss to determine how you determined superiority in theory.

Dr. Bernardin: Not in theory; no, I never said that. Just testing it. I'm using internal psychometric characteristics. They have been criticized, but using interrater agreement, which is probably not as subject to criticism as your second class of criteria in the paper. I also used halo and leniency.

Dr. Ree: It's very difficult to find a set of data that conform to Guttman's theory of his perfectly scalable items. It's easy to do when you're measuring something physical like weight. If I weigh more than 180 pounds, I weigh more than 170 pounds. But at any given time we have reversibility. If I think, for example, this color has more gray saturation than this red one and this green one today, I may change that tomorrow. I'm not sure why that is, maybe that's the question everybody's asking today. But I've yet to see any data that seems to support this work. If we look at it in an educational domain--there was a very long review article about 5 or 6 years ago on hierarchies in education that failed to find high coefficients of reproducibility among educational opinions.

Dr. Bernardin: Related to that, and I think that speaks very well to the breakdown of this, I locked at one researcher's police scales--in police work, most people just laugh at those scales. These are a disaster, you can't rate people on these scales, even though, I think, they're much better than the Landy and Farr scales in terms of their specificity. The incidents are very specific. But just to give you one example at this point to illustrate my dimensionalization problem, the way we dimensionalize things, they had a scale called "using force appropriately" and two critical incidents that were right next to each other on rating effectiveness. One was about beating up a suspect and bringing that suspect into the station after you've beaten him up, and the next incident was about firing off his or her pistol too often. That was dimensionalized as using force appropriately and scaled as such. What you're supposed to do when you're doing your rating is put your ratee in that position and rate him or her in that situation. Would the ratee do that, would the ratee be better than that, etc., next one all the way down and stop where you think it's the best example and make an X at that point. Well now, I'm not sure that those two behaviors really make that much sense. Conceptually they're using force appropriately, but if I beat somebody up, does that mean I'm going to shoot off my pistol? Maybe not. But yet if you go through the BARS procedure . . .

Dr. Mullins: That depends on how big you are, John.

Dr. Bernardin: That's really true. That is actually true. There are some L.A. reports that came out that actually relate height to the type of force that you use.

Dr. Ree: Are you saying that unidimensionality here is a problem?

Dr. Bernardin: I'm saying that using the BARS approach, we dimensionalize things that may not be legitimate dimensions; they may be illusory.

Dr. Ree: There may be techniques that are available, for example, for determining unidimensionality of scales.

Dr. Bernardin: I know, you're hinting at your paper. I agree with that. I just don't think it's a good idea, even though I have talked about the BARS approach and indorsed it in a couple of papers, to have the future rater population sit down and say, "Now are these two incidents an example of that dimension?"--because it's really feeding into their illusory correlations that may not relate to one another in the real world.

Dr. Borman: That particular problem, it seems to me though, could be a problem in actually developing the categories in the first place, the performance dimensions. It may be that a reaction to that would be to put all the incidents that fell into a couple of different dimensions back together again and redimensionalize that part of the performance domain so that the incidents would eventually get sorted back into the categories reliably. So that they make more sense when they appear together on a scale.

Dr. Bernardin: I discussed that. Regarding that qualitative cluster analysis that Campbell, Dunnette, Arvey, & Hellervik (JAP, 1973) like to use versus generating dimensions from the rater population, I like the former approach better because the behavioral groupings make more sense as opposed to having the rater population generate them. With the latter approach you get more trait oriented scales, like Landy and Farr (JSAS, 1975). Look at the Heckman scales or behavioral groupings. Look at Landy and Farr's scales. They have traits like attitude, demeanor, leadership qualities, and motivational qualities. There's a big difference there. I think it's just based on crude dimensionalizing.

Dr. Kavanagh: The point that you make about the BARS--essentially the BARS procedure involves a cluster analysis by people. And as we know, there is no reason to expect the factor structure across different populations to be the same. When we examine some of the field tests of BARS, we find they don't hold up as well as they should. But I've found, in working with BARS on a couple of occasions, that the raters accept it. The people who use it like it, and if you go back to my earlier comments, it increases their confidence in the system. I'm not belittling the psychometric evidence, but I think that's an advantage of the BARS even though it hasn't delivered as completely as promised.

Dr. Bernardin: Friedman and Cornelius (JAP, 1976) looked at rater participation and found differences. I did that but didn't get differences. Warmke and Billings (JAP, 1980) looked at that and also got differences as a function of participation. There's some therapeutic value in just participating in scale development. Maybe these people haven't even ruminated over effectiveness levels of various dimensions of performance before. Let me go on to cost effectiveness, something that Wally mentioned in terms of the rater scale development and how summated scales are probably just as good as BARS. In my studies it's probably just as costly and time consuming to develop good summated scales as it is to develop BARS, so I don't think that that, in terms of cost effectiveness, is an important argument.

Dr. Kavanagh: Maybe. But I think the cost effectiveness in terms of the indirect cost of people time in the organization may be greater.

Dr. Bernardin: Yeah, in summated scales it's probably more in line with your job to put them together, do item analysis, and that stuff, whereas all those iterative meetings for BARS would take up a little more time, probably.

Dr. Borman: At least in our Navy recruiter group, we actually found some resistance to the BARS form that we have been using for a while, and those so-called behavior summary statements that are a bit more generic and that describe the behavior depicted in a number of behavioral examples at a particular level of effectiveness turned out to be much more acceptable. I don't have any data on this but in talking with many many groups of raters it seemed to me, at least, from the content of their comments that they could handle those kinds of statements much more easily. But the main problem there is we may have had a kind of unique situation in that these guys' jobs were different really, in different parts of the country. For instance, one example is if you have a critical incident about a recruiter entering a school to do some kind of recruiting, well it turns out that some recruiters are not allowed in high schools at all. And so, obviously, a critical incident that involved that kind of behavior would just not be relevant to some proportion of the recruiters. And so this was the kind of problem this summarizing seemed to solve. But of course that would not apply to a single setting where everyone was really pretty much under the same kind of criteria.

Dr. Cascio: I have a different hypothesis for that. I developed BARS and summated scales for prison guards. This was the project I was doing last year. And one of the things that we found, and have some test data to back this up, is that these people found it far easier to deal with concrete statements than abstractions. And in dealing with the true BARS format, what they found was they'd say, "Well, I've never actually seen the guy doing this. I can't put him in that category because I've never seen him doing it." What we've found is if you take the summated type scales--well, "Here's an example of how a prison guard would perform if he were rated outstanding." We have three or four examples, all of which clustered about the 7th or 8th or 9th point on the scale. "And here's how a person would behave if he were rated average." There are examples. The summated scales seemed to work a lot better because the guards had a greater facility in dealing with the concrete than with abstractions. They could not abstract. They'd say, "Well I don't know where he fits." Simply another hypothesis.

Lt Col Ratliff: There have been some cross-culture comparisons where people have been asked to describe other people and observations made of how they do it. For example, if you ask an Englishman about a friend, he will tell you how closely he conforms to the stereotype of his job and station and where he deviates from it and gives you a pretty good detailed view of him. A Frenchman will go into the person's emotions, his tastes, whether he's a gentleman or not, honorable, etc. If you ask a Russian you get another 20-minute dissertation about his reputation, whether he's reliable in returning papers, whether he's courteous under certain conditions, etc. If you ask an American, he says, "Oh, he's a good guy."

We seem to be dealing with a series of scale dimensions here, in trying to get people to rate on behaviors that they may not consider relevant to the kind of evaluation they think ought to be made. In looking at the format problems and the scale problems that we've had, I know in this Laboratory several years ago in the Occupational Analysis Branch they were messing around with scales. They found that if you just drew a scale, titled it, and put a 1, a 10, and a 5, at equal distances, you'd get pretty good distributions on it. It didn't make any apparent difference about elaborateness and format. I probably would agree that there is a cost effectiveness factor in going up from this simple format.

I'm wondering if there's been any studies done where you've asked people to comment on other people, ad lib, without trying to structure them at all, and to see what kinds of things they say.

Dr. Kavanagh: It's interesting that you mention the cross-cultural effects. When John started talking about differences in people, synthethizers versus analyzers, one of the things I was emphasizing with Wally's paper was the whole notion of some cultural determination. If we introspect and think about the way we've grown in our culture, we're taught not to analyze people's behavior but rather to come to some global judgment. More important, I think, is what type of information do our institutions use to make decisions about people?

Lt Col Ratliff: I think that's very cogent. It may not be related to a formal model--it's probably not--there's an informal decision making process that's much more pervasive.

Dr. Kavanagh: That's the administrative side of performance appraisal. If we believe that there's a truism that people like to be evaluated and found out to be pretty good, then BARS serves that purpose in terms of being able to specifically tell them where they're good, to some extent. The global thing may do the same thing. But we know that not everybody in our organization is going to rise to he chief executive officer. So, BARS may give them at least some partial positive reinforcement.

Lt Col Ratliff: I have a deeper question. Really, what are you trying to do with BARS? Why are you rating the individual? Are you trying to measure job performance as such? What is done on the job? Or, are you trying to say, "This individual has these traits; therefore, I am certain he will do a good job?" Or "He has these traits; therefore, I will predict this about him?"

Dr. Bernardin: Are we trying to correlate traits with performance or are we trying to rate pe formance? That is lost in the BARS literature; they don't deal with the dimension. It's what I referred to earlier by Landy and Farr; they're rating on traits and they're inferring that the traits are somehow correlated with performance. The way Smith and Kendall set out in their methodology was to have performance dimensions up here, not traits, very specific kinds of categories. I don't know why but I keep thinking of using force appropriately. Okay, that's a behavioral grouping from Heckman (Technical Report, PDI, 1973). I can say, "Well, I don't like it, it's brutal or sadistic." That's a behavioral category; that is a performance measure; you rate on that performance measure. There's no inferential jump that you have to make from that trait, measuring attitude, like in the Landy and Farr case, to the performance. So as these were initially developed, they were supposed to be performance ratings on performance type dimensions.

Lt Col Ratliff: Sort of like those on OERs, APRs, and other performance rating kinds of things.

Dr. Bernardin: I'm not familiar with OER . . .

Lt Col Ratliff: Well, the OER has certain kinds of statements--judgment and cooperation . . .

Dr. Brokaw: It's an officer effectiveness report.

Dr. Bernardin: I knew what that meant but--not really; judgment is almost more in the trait camp. When Smith and Kendall developed their scales, they were talking about dimensions that were not that trait-directed. They were much more behaviorally descriptive than something like judgment or attitude or motivation. Even though other people have defined motivation in more behavioral terms and then had these critical incidents scaling them, Smith and Kendall intended that the dimension be a performance oriented kind of dimension as well.

Lt Col Ratliff: I think the problem that I may have is, if I'm talking to Dr. Mullins, or somebody else in the shop about another person and asking, "Can he do this job?" We never use any terminology that I can even infer with any sort of scale. We talk about something else. Our frame of reference about the job and our perception of the individual is somewhat different. If we wanted to make a rating scale out of the judgment that we had made--whether the person would or would not or could not perform--it would be very difficult to put it into that format. I think that's the problem that I have.

Dr. Bernardin: Are you talking about ratings of potential for another position?

Lt Col Ratliff: Yes, or, who should be assigned to a task in the unit. The parameters on which performance is measured should probably approximate those from which you predict performanc

Dr. Mullins: I think that may be part of the problem. I think if you're talking about how someone will fit into some group, you're really talking about a prediction problem. I'm not sure that you can use behavioral statements in that case because the person hasn't behaved there yet.

Lt Col Ratliff: Even in a group where the person has behaved and it's a reassignment of tasks, and a question of who can get it done the quickest, or who will do the "best" job with it.

Dr. Borman: I'm sure there's some variation in the way different people dimensionalize the job, which is what you're saying, but with the Smith and Kendall approach, originally at least, with a lot of participation by the people who are going to actually use the scales, it seems to me you'd at least be more likely to develop a set of categories or dimensions that would reflect the way people think about the job.

Lt Col Ratliff: I would grant that's true.

Dr. Borman: You know, rather than the psychologist just in the abstract coming up with traits or maybe even performance dimensions that he or she thinks are reasonable.

Lt Col Ratliff: I think part of the problem is that many of the kinds of behaviors that we can talk about in ordinary, everyday language and say get a little deeper than "he's a good ole boy." It's very difficult to verbalize in some formal way traits we can scale. You know he has certain problem solving traits, or he doesn't. We feel that he can do it because of his past experience in training or whatever, or he can't. We need somebody to assign to that job. If you have three people who are well qualified that can do it, then you have a real problem, I think, in discrimination. And, I think we have it as people trying to make that decision as well as trying to see it reflected in a scale.

Dr. Kavanagh: That is the emphasis I was putting on the binary decisions in personnel.

Let me explain a system briefly that included the two things you're talking about. Organizations use the results of our performance appraisal systems. We say, "what can this person do?" We rate him on potential or we rate him on an overall evaluation. That gives individuals very little information on how they can change or improve on their jobs.

Mike Beer and his associates developed a system focusing specifically on this issue--results versus how to improve. What they did was to develop a critical incidents technique from managers' self-reports. Then they did ipsative factor analysis, and found they could actually generate ipsative profiles on individuals. The rater could then provide feedback on strengths and weaknesses to the ratee, with this computer-generated profile. They went through the performance rating and said, "Now this is how you can change; this is what you can do." It's like a salesman, or two salesmen both getting the same number of sales, but one was cheating to get the sales. You know that's no good. So you've got to impact on the processes of behavior, and I think that is the critical part of their system. That profile, by the way, never went any further than the supervisor's desk. There was an additional one-page summary that was used for administrative purposes and rated overall evaluation. And this last rating, Colonel, that one page rating, did not go in until after the counseling session on the profile. Then the rater and ratee came back a week later and talked about the overall evaluation, promotion, and transfer. They've got an extremely well developed system, one that is focused on the concept of both administrative and feedback purposes. I think that's what the behaviorally anchored rating scale gives--an opportunity for a supervisor to come back to an individual and say, "You're performing at a fairly acceptable level but you could improve, and here are the behaviors that I think you could improve in."

Lt Col Ratliff: It's like a feedback mechanism?

Dr. Kavanagh: I think more than anything else.

Dr. Bernardin: And that's a thing that's never been tested, by the way. Cummings and Schwab (Performance in Organizations, 1973) talked about how it should facilitate improvements in performance because the feedback is so specific, but that's never been tested.

Dr. Borman: Ideally, though, in addition to what you're saying, you know it hasn't worked out this way in research but it still should provide an easier kind of process for raters, if they can match observed behavior with behavior on a scale. Ideally that sounds like a very reasonable kind of process to help raters to rate more accurately. So I would say in addition, at least that's the intent, an additional intent.

Dr. Bernardin: Let me say one more thing on my paper. I didn't bring the rater training thing up at all but I just finished a study that looked at what Wally was talking about in terms of relationships between validity and psychometric error, and it's clear that some of the stuff that we've been calling rater training is probably training response set. We can get them to make less lenient ratings and less halo effect as measured but it has nothing to do with accuracy at all.

Wally mentioned developing frames of reference or stereotypes. That's the research I'm doing now, developing stereotypes of effective workers and using that as a training device along with training on diary keeping procedures and the like, so I think that's the route to go in terms of rater training. A final thing, and this is amazing because Wally and I did not exchange these ideas, but my recommendations for a behavioral observation system are very close to what Banks is talking about in terms of content analysis. And I think that's a better approach to rating where you separate the appraisal system from the observation system. The observer merely enters the observed behaviors perhaps on a terminal and then someone else rates those behaviors on effectiveness. It obviously has problems, many, but I think it may be a better approach than the standard rating procedure.

Dr. Mullins: I was fascinated with that in your paper and I want to talk to you some more about it when you find some time.

As a side comment I'd like to mention to you something you may find interesting. We do have a contract in the works, I think it's already been let, which will construct a package for training people and then when we get that back in-house, we're going to do the training to see whether or not the rater accuracy index goes up after training or stays the same. We're going to validate it against some outside criteria. I think that's scheduled within the next 6 or 8 months.

Dr. Cascio: I think the most important aspect of John's paper which we didn't even talk about yet, which he didn't talk about, was the need to take context into consideration in developing appraisals. It's been sadly left out of a lot of behaviorally anchored rating scales and in our rush to abstract these critical incidents, many times we've abstracted the context right out of them. And that to me is a very critical part of any type of evaluation.

What I've heard this morning, so far, is a plea to develop better observation systems. We know very little about how to train raters, how to train observers. We know very little about that. One of the things that I think we'll all stress in training observers is the need to take context into consideration. Because when we strip the context away from it, we're left very frequently with behaviorally sterile descriptions, and if we ask people to look at a behaviorally anchored rating scale where we have these behaviorally sterile descriptions that are on the scale, we sometimes lose the richness that observation provides. We need to take contexts into consideration. I liked your term reality-based behavioral systems that take context into consideration.

I think you can't deny the fact that global criteria with all their defects, all their known defects, still produce validities that are as high or higher than any other rating system that we've developed. I suspect that one of the reasons for that is because we take the richness and the context into consideration when we derive those global ratings.

Now that's entirely separate from consideration based on fuzzy criteria, as in Moody versus Albemarle where the Supreme Court spoke very clearly and very specifically about the fact that paired comparison ratings in which one individual was paired against another and supervisors were just instructed to say which one of these is better, which one of the two is better. And those kinds of systems, as you know, were struck down because we didn't know what criteria the raters were using in saying that individual A was better than individual B. Nevertheless, there's a lot of work that's been done at the Industrial Relations Center at the University of Chicago, Melanie Behr and her associates, over a period of years, which has shown again and again that these paired comparisons are effective as performance appraisal systems if we're only looking at them in terms of administrative decisions. Of course, what they don't give you is the behavioral specificity which you could use in personnel development. We can't deny that fact.

I also like, just to jump off on one other thing, I also like your idea of the theory of conceptual likenesses. I think one of the things you didn't bring out in the paper, which we can talk about, is that that's empirically testable. We can get at that if we can perhaps get at the kinds of theories that people subscribe to as to what relates to what, the basic idea being that there's a need to inject predictability and order into the world, into the chaos, the variation that surrounds us. And this is why schemes like McGregor's theory-X, theory-Y have been so effective, because they're easy to understand and they enable people to pigeon-hole individuals into one of the two categories. It's very simple to understand and fulfills this need for predictability and this need to see order in the world around us. In this theory of conceptual likenesses, each individual's idiosyncratic ideas about what relates to what are empirically testable. I think we can get a handle on this and then perhaps ultimately relate that to the performance appraisals that they award their subordinates. It's definitely empirical. As a matter of fact I want to test it empirically. I think it's something we need to latch onto.

So to summarize then what I've said, two things, two main points, one of which is that we often lose behavioral richness when we strip context away from content, exactly what people did. And secondly that this theory of conceptual likenesses may explain a lot of the halo that we see in ratings but it is empirically testable. We can get at that and determine what kinds of schemes people subscribe to as to what relates to what.

Dr. Kavanagh: I have a question that I am not sure of the answer. I'm concerned about what sort of systems are going to look good in the courts.

Dr. Cascio: I've been involved in a case recently where that was an issue, and at least in my experience the key thing, and I go back to perhaps one of the earliest cases dealing with performance appraisals as a legal issue, Wade versus Mississippi Cooperative Extension Service, where the performance appraisal was used as a predictor of later performance and therefore legally is regarded as a test. Anything used as a basis for personnel decisions. .And the critical criteria that were brought out in that case, and the one which I have recently been involved in, is, "Can we specify the basis on which people are making decisions?"

That seems to be the bottom line, at least as far as what judges are looking for. Take these paired comparison, global ratings, and you can get into a lot of trouble if they produce an adverse impact. Of course if they don't produce an adverse impact it's irrelevant because you don't even fall under the Guidelines. But if they do, and 90% of them do, then you've got to show the basis on which people are making decisions. And more importantly, show that they were not making decisions on the basis of race or sex or any other impermissible factor.

Dr. Bernardin: Fred, are you going to bring that up, what we talked about? Because just reading Cecil's paper and one statement "the simple difference c. elevation at the beginning observation point means a considerable difference in developmental level later" . . . . It seems, at least in your presentation, that you expect racial differences on any performance measure that's a valid measure, based on your presentation. Meaning we will probably get adverse impact with a valid system, so how can we defend it in courts is really a crucial issue.

Dr. Cascio: What cost the Mississippi Extension Service that case and why I think we won our case locally in Miami was because we had done a very very comprehensive and thorough job analysis before we even developed the performance appraisal system, and this of course is one of the major complaints with respect to lots of performance appraisals that are done--they're not job relevant. And how do you show job relevance the way you do with job analysis? If you don't do that, if you just jump and put the cart before the horse and try to develop the performance appraisal system first, you can get into a lot of trouble, especially if there's an adverse impact, and there will be.

Dr. Muckler: I'd like to pick up on this thing that Wayne said, what criteria the rater's using, then switch back to what John was talking about on the clarification statement. Now that's a simple thing and it ought to be reasonable and it ought to be a nice thing to do to put little statements out there to help people. But everytime we try to do that, it's just opening Pandora's box. We sit down, we think up little statements, and I think we make one mistake. We test it on our colleagues. If we didn't do that, we'd be all right. But invariably, you put down a simple example and somebody says, "Well I don't think that's so good," or "I don't think that's so bad." And if you explore this, the next thing you know you've got a Rorschach going. But it's really interesting because then you start asking, "What do you think is good performance?" and "What do you think is bad performance?" and the next thing you know you're getting into what the guy really feels is good and what is bad. And there are three things that we find when you do this. One is that supervisors in comparable positions disagree very fundamentally about what is good and bad performance. The second is that they're ambiguous about it. It's very difficult to verbalize it. And the third thing is it's threatening to them. We were doing this in one case and by accident one of our very top managers got involved and we sort of began using a psychoanalytical approach on him beause we really wanted to find out what he thought was good and what he thought was bad. And I think we got a pretty good idea, and I think we were pretty dismayed about it. I'm not sure that we can ever get around this. I really got to know what the rater thinks is good and what is bad.

Dr. Mullins: I'd like to make one comment. Again, when you see our system later on I think we've managed to avoid that issue pretty well because we started off with the assumption that two people with exactly the same position description may very legitimately do quite different jobs. We started off then with that assumption that you can't take standard functions to rate the people on because some of them might not even apply. Some research psychologists might do so and so, some might do something else, and they might not even overlap. Well, I'm not going to go into the entire system. You'll see it later. But there is a way that that can at least be alleviated. You'll see it later on. But that's an interesting point. I would like to caution against feeling that just because different supervisors disagreed with those statements, that necessarily that statement's bad. It may be that in this job situation this is important and in that job situation that is important although both the two people being rated may be classified exactly the same.

Lt Col Ratliff: You might have such enormous supervisor variability, as I interpret what you're saying here, with so many built-in little hidden perceptions, that the more you learn about the process, the more dismayed you're going to be anyhow, and it may be well not to be too precise.

Dr. Bernardin: That's one of the problems with the BARS approach. In those iterative steps, we're looking for consensual agreement on incidents and on dimensions, and there may be a good deal of disagreement that is expressed or not expressed and that may be the crux of the rating, that hidden stuff. You've got these rating scales with all these terrific anchoring incidents that have no relevance to the actual rating process that an individual rater is making because the stuff they disagree on doesn't survive the process.

Dr. Muckler: And in one case where we attempted to get a consensus in this kind of situation the atmosphere became very cool and very hostile. And I'm not sure it was possibe to resolve the differences.

Dr. Mullins: Okay, Wayne, I believe the podium is all yours.

CHAPTER 4

HUMAN ASSESSMENT: WHERE WE ARE AND WHERE WE ARE GOING

Wayne F. Cascio
School of Business and Organizational Sciences
Florida International University

Invited paper prepared for the U.S. Air Force Human Resources Laboratory Conference on Human Assessment, Brooks AFB, San Antonio, Texas, March 1979.

Human Assessment: Where We Are and Where We Are Going

Physical and psychological variability is all around us. As behavioral scientists our goal is to describe this variability and, through laws and theories, to understand it, to explain it, and to predict it. Measurement (or assessment) is one of the tools that helps us along the path to this goal. However we must first understand the logic (the why) of measurement before the how--that is, measurement techniques--becomes more meaningful. So in order to set the stage for some of the points to be brought out later, let's begin by considering assessment briefly from four perspectives--what, why, how, and how good.

Q.: What is our objective in assessment?

A.: To measure individual differences in physical and psychological characteristics in order to make inferences about the relative standing of each individual on the physical or psychological characteristic(s) in question.

Q.: Why? For what purpose?

A.: To make decisions about individuals. In selection, the decision is whether to accept or reject an applicant; in placement, which alternative course of action to pursue; in diagnosis, which remedial treatment is called for; in hypothesis testing, the accuracy of the theoretical formulation; in hypothesis building, which additional testing or other information is needed; and in evaluation, what score to assign to an individual or procedure (Brown, 1976).

Q.: How? What processes are used in assessment?

A.: Let us consider selection and performance appraisal as examples. In selection we attempt to predict relative job behavior effectiveness on the basis of available information. However, this is a two-step process: data collection (e.g., through written tests, interviews, relevant background information), and data combination in such a way as to enable the decision maker to minimize predictive error in forecasting job performance (Wiggins, 1973). In performance appraisal (the systematic assessment of strengths and weaknesses within and between employees), a two-step process is also involved: Observation and evaluation of what is observed (Guion, 1965). Notice the parallel processes of assessment in selection and performance appraisal. The foundation in both cases is data collection (observation simply being a method of collecting data). The second step is to combine the data in such a way as to formulate an evaluation, a prediction, or both.

Q.: How good? What criteria can we use to evaluate the goodness of psychological measures?

A.: Since we are attempting to assess and predict status under a variety of conditions, perhaps the most appropriate criterion for evaluating psychological measures is in terms of their social utility (Comrey, 1950, 1951). The important question is not whether the psychological measures as used in a particular context are accurate or inaccurate, but rather how their predictive efficiency compares with that of other available procedures and techniques.

Thirty Years of Progress--And Lack of Progress--in Assessment

During the last three decades the field of human assessment has made some notable advances, but some of our shortcomings have been notable as well. In general, advances have occurred in three major areas--standardization, quantification, and understanding. In stressing the importance of standardization we recognize the importance of specifying first, as rigorously as possible (e.g., through comprehensive job analysis), the content domains of interest, so that our inferences about individuals will be limited to those domains. Thus we recognize the wisdom of ensuring that our selection procedures (tests, job samples, assessment centers), our training programs, and our performance appraisal systems, are firmly grounded in job-relevant patterns of behavior. We have also come to put great stress on ensuring the reliability and validity of our methods of assessment. This is as it should be, for only those procedures which can pass "psychometric muster" have the potential for advancing our understanding, for enabling us to make sense out of the numbers we attach to various levels of performance, and ultimately for leading to the development of testable theories of the behavior of men and women at work.

On the other hand, there have been important aspects of human assessment that we virtually have ignored--these are our deficiencies to date, and they also should be made public, so that we can work to improve them. Consider two important requirements for any assessment procedure--relevance and acceptability. To date we have put considerably more emphasis on the former than the latter. In many, if not most, assessment programs (those dealing with selection as well as performance appraisal) we have not put enough effort into garnering the support and participation of those who will use our procedures. The accent has been more on technical soundness than on the attitudinal and interpersonal components of assessment programs.

Psychometrics and organization development have not been given equal weight in either development or implementation.

Consider the typical selection program for example (educational or occupational). Assessors diligently attempt to keep answers to many of their tests a secret, lest people practice and learn how to do better on them or fake high scores. However, faking is impossible as long as a person is performing actual criterion behavior--the computer programmer applicant actually writing a program, the proofreader applicant taking reading comprehension and spelling tests. Faking only becomes possible when test behavior and criterion behavior are not directly related. Under these circumstances, when applicants know very little about what will be tested, how they should prepare, or how the test will be scored, it is no wonder that they are suspicious, and in some cases, devious and untrue about themselves. The same pattern of behavior was found in the forced-choice performance appraisal checklists which were so popular following World War II. Near total emphasis was placed on technical soundness, and near total de-emphasis was placed on the acceptability of the system to those who used it. Under these circumstances, according to McClelland (1973), we are playing power games with people over the secrecy of answers and pretending knowledge of what lies behind correlations, which may not in fact exist. On top of all of this, only rarely are individuals told about the relative strengths and limitations of the assessment procedures. This kind of frustration has led to pressure for truth-in-testing legislation, such as that passed in California in Septemer 1978 and pending in Texas, Indiana, and New York (Testing Digest, 1979), which forces assessors to be accountable to assessees.

How much simpler it is to enlist the active support and cooperation of the test taker or subordinate by making explicit exactly what the criterion behavior is that will be tested or appraised. In contrast to promoting secrecy in selection and performance appraisal, it is my opinion that we should be promoting more openness, so that we can say, "This is what competence means in this situation; this is what you must be able to do in order to perform competently." When teacher and pupil, boss and subordinate, assessor and assessee can collaborate openly in trying to improve performance, when how to pass the test or how to improve performance is public knowledge, when both the assessment technique and its practical utility are understood, then we can expect to find the active participation and support for appraisal programs that is so sorely needed in the field of assessment. Recent evidence reinforces the merit of this approach (Cascio & Phillips, 1979).

Research Needs of the Future

Before human performance can be studied and better understood, four basic problems must be dealt with (Ronan & Prien, 1966, 1971).

These are the problems of reliability of performance, reliability of performance observatio., dimensionality of performance, and modification of performance by situational characteristics. Although I have discussed each of these issues in detail elsewhere (Cascio, 1978), I would like to give special emphasis to the problems of reliability of performance and reliability of performance observation, for they have been neglected to date.

The few studies available in the literature on performance reliability (e.g., Klemmer & Lockhead, 1962; Owens, 1942, Rothe & Nye, 1958) indicate consistently that intra-individual differences in performance are significant. Thorndike (1949) identified two sources of such variability--intrinsic unreliability (due to personal inconsistency in performance) and extrinsic unreliability (due to sources of variability which are external to job demands or individual behavior). There is almost no extant research which has attempted to identify these two sources of performance unreliability in an operational setting, and to measure the magnitude of their relative effects. While sources of extrinsic unreliability (e.g., machine downtime, delays in supplies or information) can be controlled experimentally or statistically (see for example, Cravens & Woodruff, 1973), this is decidedly not the case with the sources of intrinsic unreliability. In the case of intrinsic unreliability classical estimates of reliability (the correlation of group absolute performance levels measured at Time 1 and again at Time 2), not only may be inappropriate, but actually may serve to obscure further the real issue. Thus if performance variability is as much a characteristic of an individual as an aptitude or a personality trait, then it is possible that variability itself may function as a useful predictor or criterion of motivation to perform the job. This is all fertile ground for research, and if we are to enhance our understanding and prediction of human behavior further, it must be tilled.

A second major area of concern, and one which actually takes precedence over the first, is the reliability of job performance observation. This issue is crucial in assessment since all evaluations of performance ultimately depend on observation of one sort or another. In fact the study of performance reliability only becomes possible when the reliability of judging performance is adequate (Ryans & Fredericksen, 1951). What little research there is to date on this question (e.g., Borman, 1974; Bray & Campbell, 1968;) indicates that different observer perspectives (e.g., supervisors, trainees, subordinates, independent field auditors) or methods of observing performance may lead to markedly different conclusions. While these studies have shown quite clearly that the problem exists, there is almost no information on how the reliability of judging performance can be improved. Thus in an extensive review of the past 25 years of research literature on training observers of behavior, Spool (1978) concluded:

> The state of the art in training observers of behavior appears to be in its infant stage. To date not much research has been conducted and of the research that does exist, most [studies contain] serious methodological flaws, and extremely few are comparative or systematic in nature. As a result very little is known about which training approach is most effective in increasing accuracy of observation and which components of training design contribute most to the overall effectiveness of the training program (p. 883).

That is a sad commentary on the state of the art in training observers, and in view of the dearth of knowledge in this area it is no surprise that the reliability of job performance observation is so nettlesome. However, Goldstein and Sorcher's (1974) "applied learning" approach to training may be a useful way to begin designing an obvserver training program. This is a 4-step procedure: modeling, role playing (practice), social reinforcement (feedback), and transfer of training. Unfortunately none of the studies reviewed by Spool (1978) involved modeling "how to observe"; the studies only presented examples of relevant behaviors and the observation instrument. Moreover, no research has been done which attempts to assess the relative contribution of each component of the applied learning model to the effectiveness of a total observer training program.

One final issue deserves mention. Throughout this paper emphasis has been placed on the fact that assessment programs include measurement issues as well as attitudinal and behavioral issues, and that we have typically placed more emphasis on relevance than on acceptability in our assessment programs. Yet if performance is truly a function of ability and motivation, then we must do all that we can to insure that all involved in assessment (assessees and assessors) are motivated to perform. It may well be (and this is empirically testable) that raters who are more involved and more interested in the task (because they were participants in a training program, or because they were deeply involved at all stages of development of an assessment system) will make more careful and more accurate ratings.

In a wider context we are concerned with developing decision systems. From this perspective, measurement and prediction are simply technical components of a system designed to make decisions about individuals. Since some degree of error is inevitable in all personnel decisions, the crucial question to be answered in regard to each assessment method is whether the use of the method results in less human, social, and organizational cost than is now being paid for these errors. Answers to that question can result in a wiser, fuller utilization of our human resources.

References

Borman, W.C. The rating of individuals in organizations: An alternate approach. *Organizational Behavior and Human Performance*, 1974, *12*, 105-124.

Bray, D.W., & Campbell, R.J. Selection of salesmen by means of an assessment center. *Journal of Applied Psychology*, 1968, *52*, 36-41.

Brown, F.G. *Principles of educational and psychological testing* (2nd ed.). New York: Holt, Rinehart, & Winston, 1976.

Cascio, W.F. *Applied psychology in personnel management*. Reston, VA: Reston, 1978.

Cascio, W.F., & Phillips, N. Performance testing: A rose among thorns? *Personnel Psychology*, 1979, *32*, 751-766.

Comrey, A.L. An operational approach to some problems in psychological measurement. *Psychological Review*, 1950, *57*, 217-228.

Comrey, A.L. Mental testing and the logic of measurement. *Educational and Psychological Measurement*, 1951, *11*, 323-334.

Cravens, D.W., & Woodruff, R.B. An approach for determining criteria of sales performance. *Journal of Applied Psychology*, 1973, *57*, 242-247.

Goldstein, A.P., & Sorcher, M. *Changing supervisor behavior*. New York: Pergamon, 1974.

Guion, R.M. *Personnel Testing*. New York: McGraw-Hill, 1965.

Klemmer, E.T., & Lockhead, G.R. Productivity and errors in two keying tasks: A field study. *Journal of Applied Psychology*, 1962, *46*, 401-408.

McClelland, D.C. Testing for competence rather than for "intelligence." *American Psychologist*, 1973, *28*, 1-14.

Owens, W.A. Intra-individual differences versus inter-individual differences in motor skills. *Educational and Psychological Measurement*, 1942, *2*, 301-314.

Ronan, W.W., & Prien, E.P. *Toward a criterion theory: A review and analysis of research and opinion*. Greensboro, NC: The Richardson Foundation, 1966.

Ronan, W.W., & Prien, E.P. *Perspectives on the measurement of human performance*. New York: Appleton-Century-Crofts, 1971.

Rothe, H.F., & Nye, C.T. Output rates among coil winders. *Journal of Applied Psychology*, 1958, *42*, 182-186.

Ryans, D.G., & Fredericksen, N. Performance tests of educational achievement. In E.F. Lindquist (Ed.), *Educational measurement*. Washington, DC: American Council on Education, 1951.

Spool, M.D. Training programs for observers of behavior: A review, *Personnel Psychology*, 1978, *31*, 853-888.

Testing Digest. A small victory in California: Truth in testing, 1979, *1*, 1.

Thorndike, R.L. *Personnel selection: Test and measurement techniques*. New York: Wiley, 1949.

Wiggins, J.S. *Personality and prediction: Principles of personality assessment*. Reading, MA: Addison-Wesley, 1973.

Dr. Cascio: I think before I start I'd like to share an observation with you and that is in terms of how you set this 3-day workshop up. Personally I'm just delighted that you've given us the opportunity today to throw out some ideas which are very theoretical. A lot of them are theoretical ideas, but I think that this is extremely useful in being able to stand back tomorrow and the next day and take a look at what we've said and how it fits into the ultimate system that's developed. And so I regarded my mission here, at least as far as this first session, as to plant seeds and to throw out some ideas that may or may not pan out in the next couple of days.

The first thing that occurs to me, having spent the last few years intensively involved in developing appraisal systems and implementing appraisal systems in organizations, which is more difficult, I think at the risk of drastically oversimplifying things, I would like to say and get on the record that I believe that performance appraisal is as much an attitude as it is a method. And I think that our biggest shortcoming to date is (to jump ahead with something that I tried to bring out in my paper)--perhaps our biggest shortcoming to date is--that we have not given psychometrics and organizational development equal weight either in developing appraisal systems or in implementing them.

We know that there are two basic requirements for any performance appraisal system, namely relevance and acceptability, and we've devoted considerable attention to relevance, to making sure that whatever it is that we're rating is job related, is an important aspect of job performance. We have not devoted nearly as much time to insuring acceptability by those who will use the system.

I have a hunch that one of the reasons why we fail to see significant differences in formats is because we fail to control for the amount of acceptability of the performance appraisal system in the organization where it is being applied. I believe that if raters are uncooperative or untrained, I don't care what kind of a system you have, I don't care how complex the format is, it's not going to work. Years ago with forced-choice systems we showed that if people want to beat the system, they'll beat it. John talked about that this morning. No matter what kind of system we come up with people will find ways to beat it unless they believe that it's important.

And so the question that came out this morning which I think Mickey brought up was (I'm going to paraphrase what you said), maybe we impose our own value systems on those who are doing performance appraisals. We think it's important that they come up with a valid and accurate assessment of how their subordinates are doing. Maybe they don't want to come up with an accurate assessment. Maybe organizationally the political climate is such that they can't afford to come up with an accurate assessment. And that's something that we need to keep in mind.

And so to jump ahead to a second point, one thing I might ask is how much time and effort and money do you want to spend in developing a system? Lots of times we impose our own value systems on organizations who are going to use these performance appraisal systems. We think it's important that the appraisals be accurate. We think it's important that the raters be trained. I've sat through rater training sessions where the raters could care less and they're just filling in time because they've been ordered to be there. Under those circumstances we're wasting our time. And so I think the major point is that we need to devote just as much time and just as much resources to gaining acceptability of our systems as we do to insuring that they are relevant and psychometrically sound.

In keeping with that idea, I might point out that a major deficiency as I've seen it in not only appraisals but also in selecting--you asked us to talk about measurement issues in general--is the secrecy that surrounds appraisal systems. Not only tests, but also lots of performance appraisal systems. I think Fred Muckler will bring this out later, that frequently subordinates don't know what the manager considers important behavior, don't know what is considered effective and ineffective behavior. If you look at the typical testing program, Civil Service agencies now are being deluged with complaints from examinees challenging the relevance of exams, arguing over various alternatives, and there's a great deal of secrecy that surrounds exams.

You may be aware of a case decided last week by the Supreme Court. A union was challenging an aptitude test, and the union wanted to get hold of the aptitude test and actually look at the test questions and look at the correct answers and the company said no, and so the union backed off and said, "We'll hire an independent industrial psychologist and you show him the test. We won't even look at it." And the company still said no, and eventually it wound its way up to the Supreme Court, and the Supreme Court sided with the company and they said the company has a right not to turn over these aptitude tests and not to turn over the questions and the answers if they choose not to do so. That's going to have tremendous impacts, I think, on organizational climate issues.

The new look in appraisal, I think, that David McClelland was foreshadowing back in 1973 in a well known American Psychologist article was to promote more openness in testing, promote more openness in appraisal, be very explicit.

I've just finished a very comprehensive study on performance testing for all kinds and all levels of jobs for an entire city, the city of Miami Beach. One of the things we did was to be very very explicit with the people taking these performance tests as well as with the raters, specifying exactly what it was that should be rated as effective behavior. We said to the ratees, this is what it means to be competent, this is what you must be able to do. They knew ahead of time, coming in to the performance test. This is what they had to be able to do. There's nothing hidden, nothing up the sleeve. After 12 months of running these kinds of tests, we compared the results of these tests with the results of the traditional Civil Service approach, the written test questions, in terms of how examinees felt about it, their attitudes in terms of the perceived fairness of the procedures, and in terms of the number of grievances that were filed, complaints about the tests. And to say that there were significant differences is an understatement because in 12 months of testing using the performance tests we were getting an average of less than 1% complaints per month. With the written tests we were averaging almost 10% complaints of those who took the tests. In perceived fairness there was no comparison. People felt that the performance testing was very fair. Perhaps more importantly, they understood what competence meant and they understood that if they did poorly, there were no complaints because they felt that it was fair.

I see that in our typical Civil Service approach to written aptitude testing we sometimes take exactly the opposite tack. We're very secretive about what it is we're going to test for and what kinds of questions people can expect to run into. And so we can't expect then to find overwhelming levels of cooperation on the part of the examinees.

I think the same thing is often true with performance appraisal systems, very sadly. People just simply don't know what is effective behavior and what is ineffective behavior. You pointed out this morning that you sat in with one of the top administrators anu were very dismayed because he seemed to have a different set of values than everybody else with respect to what's effective and ineffective.

Lt Col Ratliff: While you're on this topic, we have had some very intensive debates at the Air Staff level over the fact that private companies always have secret appraisals; i.e., they have a chance to be more frank about their people. Thus, they really know who the good guys are. Yet when I'm looking for objective evidence from these companies that this is the case, even the people involved in the personnel systems snicker at the idea that their ratings are more objective. If I'm right, I hear you defending openness. Are there any other studies on openness that you can think of? To us, this is a very important issue as we're stuck with it.

Dr. Cascio: The Personnel Psychology article, the system at the glass company that was described . . .

Dr. Kavanagh: They had independent psychologists come in and evaluate the system. It is also a pretty open system in terms of what's going to be done with the data on employees. They also did intensive interviews with people where the system had been implemented and found that it was acceptable. In some of the divisions of the company, it had been mandated by management. The vice-presidents of the division just said you will have this system. And even in those divisions the outside investigators found highly positive responses. The conclusion of their study was that this system is so good that you can impose it without elaborate implementation efforts.

Dr. Mullins. I have a question along that line. Will your Miami City study be published?

Dr. Cascio: I hope so.

Dr. Mullins: Can I get a copy when it comes out?

Dr. Cascio: Sure. I'll send you one in advance. Actually, it's going through its second round of reviews now. It's the first effort I know of that's evaluated an entire program of performance testing. It's performance testing, not performance appraisal. But I'm going to tell you about an appraisal system too that illustrates the same problem with the same possible results.

Last summer I was called in by a bus company to evaluate a performance appraisal system that they had working. They had already had some evaluations on that appraisal system and they were uniformly negative. These were people who were brought in to look at it--What do you like? What don't you like? I'm going to tell you about the system, and you're going to say, "Wow, that really sounds simple," and it is simple.

It works like this. The bus drivers are rated quarterly. Every quarter each bus driver starts off with 100 points. Each bus driver is given a sheet, and on the sheet there is a list of about 15 different types of behaviors that a bus driver could do and they were all negative behaviors--gets into a preventable accident, fails to stop for someone who is waving for the bus, shows up late (it's very important that they be punctual because they've got to keep the schedule), is discourteous to a passenger, etc. Each one of these violations carries a certain number of points that are deducted every time a person violates that rule. So at the end of each quarter your performance appraisal is simply whatever points are left out of a hundred. Each year, the company takes an average of four quarterly ratings, and that constitutes the driver's rating for the year.

A lot of people had objected to that and I'm sure I can anticipate a lot of your objections saying that, "It's all negative. Where's the positive in it?" Now there's no reward for good performance, just punishment for bad performance. Well, that is true; we need to be aware of that. But just as important was the fact that when I went out and interviewed not only drivers, but supervisors, and most importantly union members, the stewards, they endorsed it 100%. They liked the system. It was very open; they said, "Well, when a man is graded down, he knows what he is graded down on. We like this system because it's out in the open and there's no two ways about it. There's no ifs, ands, or buts about it."

The system is also working. It works for purposes of merit pay, for bonus increases, for disciplinary reasons, and for development purposes. There were modifications that we ultimately made to that system but the basic structure of it hasn't changed. They're still using a point system which is very simple. You're dealing with people who have eighth and ninth grade educations and the thing works because it's open and above board.

There's an example of a performance appraisal system where it has the acceptability of those who are going to use the system, and lo and behold, it works.

Lt Col Ratliff: The other thing you commented on was acceptability, and I wondered if you had any measures of that because we have a survey going out asking about acceptability of various measures, and we hope to have baselines if we are able to give it again.

Dr. Cascio: Okay, in this particular incidence I don't, because I was called in after the fact. This thing has been running for awhile and I wasn't in on the ground floor when it was being introduced, but I can tell you this, that I was impressed at how strongly the union, supervisors, and subordinates defended this system and didn't want to see a lot of changes in it as a result of the reports that had been written by other consultants who wanted to inject changes in the system. Maybe some of that was resistance to change, I don't know, but they sure endorsed it.

Lt Col Ratliff: So with this system, you did not have an outstanding. What you really had was various degrees of negative and if you were clean, you were a good guy.

Dr. Cascio: Oh, you got 100%. You had 100 points at the end of the quarter. But if you had a preventable accident, for example, it cost you 25 points, so you were down to 75 right there.

Dr. Kavanagh: Let me continue the Colonel's train of thought. Something I've been struggling with is this whole concept of acceptability and how to evaluate it. I got involved late in the glass company system. I came in as an evaluator. While talking to Mike Beer, I asked him "How are you evaluating acceptability in this system?" And he said to me, "Client satisfaction." I asked him what he meant. He said "Managers like it." I said, "Where's the data? That's not a very scientific way to approach this." To truly test acceptability requires some strong manipulations in the field that we typically are not capable of doing or, at least, some data collection.

Lt Col Ratliff: All we can get is survey data taken over time. It's useful, but I don't know how strong it really is.

Dr. Kavanagh: One of the ideas I have been considering is to collect time series baseline objective and self-report evidence, in the typical pre- and post-design. If your present system has ills in it of some sort, and they're identifiable, they should be measured. Perhaps the measures would be changes in hard data such as turnover and absenteeism as well as changes in survey data. This is the difficulty one gets into in terms of the evaluation of system-level implementation unless you can do something like taking Wally's work with rating true scores. If you could say that within any organization there is a set of levels of organizational functioning such that, to be a good organization, morale should be at this level, turnover should be at another level. Then it would be easy to evaluate a system. If we look at the traditional indices that we have used, and then we implement the system, the indices should move in a significant way. Other than that, I don't know what one might do to evaluate at the total system level.

Lt Col Ratliff: Where we've run studies, we've done some very intensive post-study interviews, surveys, etc., for feedback. What we find is that the workers are happy with the old system; everybody gets good marks. The supervisors know the whole thing is a sham; therefore, they say no system is going to work because they're going to game it.

Dr. Kavanagh: I have another suggestion: use aggregated data analysis with large surveys. Also, time-sample selected individuals and do intensive case studies. Look at changes in those individuals on certain important variables at various levels in the organization. I feel that's legitimate data for evaluation.

Dr. Cascio: Those are knotty problems and we can talk some more about them in the next 2 days. As I said, I just want to start to plant some seeds about the research needs of the future.

There are two big ones that I see in performance appraisal. One of them has to do with reliability of performance and the second is reliability of performance observation. Reliability of performance--we need to be able to partition intrinsic unreliability from extrinsic unreliability. Intrinsic unreliability stems from personal variability in work, personal inconsistency. But extrinsic unreliability stems from conditions beyond an individual's control, machine downtime being a prime example, or as we were talking about sales performance, some territories are more lucrative than others. They have more accounts. A company spends more on advertising in one particular territory than in another.

There are very few studies in the literature, Cravens and Woodruff in 1973 being one of the few, where they investigated sales performance, took all of these extrinsic factors, if you will, all of those conditions beyond the individual's control, and cast them into a regression format.

We need to separate extrinsic unreliability from intrinsic unreliability. And what do we do with intrinsic unreliability? Intrinsic unreliability or personal inconsistency in job performance I believe is an excellent indicator of motivation on the job. And, lo and behold, the study you pulled out this morning from Organizational Behavior and Human Performance (OBHP) seems to reenforce that very nicely.

Dr. Kavanagh: High variability workers, in terms of performance, were judged to be more able and less motivated than low variability workers.

Lt Col Ratliff: Is that "able" in terms of actual job skills or potential in the cognitive dimensions?

Dr. Kavanagh: Ability to carry out a prescribed job, which happened to be an experimental lab study of marble sorting.

Dr. Cascio: They're less motivated. The higher variability types were less motivated on the job. It's an excellent indicator. An unobtrusive indicator too.

Now a prior problem is the reliability of judging performance because since all appraisals of performance ultimately depend on observations, then reliability in judging performance becomes extremely important, and this has to do with the problem that we've all been talking around and around this morning, and that is training observers. We know very little about how to train observers to improve the reliability of the process. One approach that I've been using in doing some job analysis involving observations of fire fighters, but is in. . .

Lt Col Ratliff: I hate to keep interrupting, but this has come up several times this morning; i.e., we don't know how to train raters to rate. We don't know what to have them rate. Is this really saying the supervisors don't know what to look for?

Dr. Cascio: Some don't. Wally has found that there seem to be individual differences in the ability to make accurate observations. And that's something that is worth pursuing, I think. On the other hand, we don't know how to train observers of behavior to increase the reliability of what they see. That's the root of the problem. How do we do it? We don't know. There's very little information in the literature about that.

Lt Col Ratliff: You're really saying that the observer unreliability is probably a greater component than performer unreliability.

Dr. Cascio: Sure. You can't measure performance unreliability when you can't get at it. And you can't get at it until you solve this problem of observer unreliability because the observers are the ones making the judgments. They're the performance appraisers, if you will. So how do we get at that? An approach that I've been trying experimentally is to develop a very detailed observation format using functional job analysis; that is, each task is analyzed in relation to data, people, and things. The same method that was used to develop the Dictionary of Occupational Titles--What did the worker do? To whom? What implements did he use? And all of that.

Lt Col Ratliff. I think that's very good. I haven't seen anybody use that in a long time.

Dr. Cascio: Well, I'll have to send you a copy of the rating form that I'm using to rate tasks. All these people do is simply describe what is done. I have two observers riding around on fire trucks and spending time at the fire house and they're making their observations independently. I don't have any final data on this yet because we just started it a couple of weeks ago; but ultimately what we want to be able to do is to assess the reliability of their observations using this very structured format for making those observations. That's one approach, but we can't expect supervisors to be running around with this kind of structured rating form.

John pointed out that the diary keeping method seemed to have some promise. I guess today all we're doing is sort of throwing out possibilities, ways of improving the reliability of observations.

Another approach is the applied learning model, Goldstein and Sorcher's method of modeling effective observer behavior. As you know, there are four steps involved in that: actually watching a model perform; secondly, role playing, that of practicing it; thirdly, getting feedback on how well or how poorly you did; and then fourthly, the organizational reinforcement for that. The transfer to the job. Again, there are no studies that have tried to apply that approach to increasing the reliability of observer behavior but it's an alternative.

Dr. Borman: Have you thought about the first step in relation to training raters? I've puzzled over that. What kind of situation do you set up for the person who's the learner? The learner is supposed to observe a successful rater, essentially, but I'm just not sure how you do that. I mean, for instance, certainly you would not videotape someone who's making a rating and say this is a successful rater, because the learner can't learn anything from how the person is holding the pen or the pencil.

Dr. Cascio: I'm not sure that I would do it on a videotape. I think what I would do and what I've done with my raters, these people I have doing the observations, is I've made sure that they understand some of the research results from the Life Insurance Agency Management Association on interviewing, the one I talked about this morning where people who watched the videotape of the interview and took notes were a lot more accurate in reporting what took place than those who simply sat back and trusted their memories. I've spent a good bit of time with my observers getting acceptability, getting them to believe in the approach of making very structured observations, and I believe that that's half the battle right there, getting them to want to do it, because if they don't, then we're wasting time.

Lt Col Ratliff: What you're basically saying is that you believe that in making your raters emulate the good raters, that in terms of behaviors they will go ahead and make better judgments.

Dr. Cascio: At this point we're just doing job analysis, and we're just trying to record what it is that these fire fighters do. What I'm interested in finding out is, "Does one observer see that task the same as the other does?" And if we can't even get agreement on that, then we have no hope for ultimately being able to improve performance appraisal systems. It's a very basic kind of approach.

Dr. Bernardin: You recommend the same thing Spool (Personnel Psychology, 1978) recommends in terms of Goldstein and Sorcher's modeling approach. I just think that there was too big a gap there between the contrived situation of modeling (we're talking about modeling a good rater) and the real rating situation, and I just don't think that that approach works; that's why I recommended a more experiential approach.

The other part I really buy. I did two diary keeping studies and I think they're good examples of how it can work and how it can't work. The first study, with very detailed follow-up, almost surveillance of their diaries, making sure they were keeping them and counting the number, and whether they were maintaining critical incidents or not just halo kinds of descriptions, yielded excellent psychometric characteristics, including good interrater agreement among people using the diary approach. The second study, which developed kind of serendipitously because I had an incompetent assistant who wasn't using this surveillance approach and kind of just let the people go, produced diaries which were terrible. The critical incidents were terrible and they were recorded with 3-week gaps in between incidents, and basically terrible. And the results were subsequently terrible ratings.

So I think the diary system can work as long as you make it a point to the supervisors that it's an important job function. It's not to be done 2 days before you do the rating. It's to be done maybe every day or every week or whatever; then it will work. Then maybe, when you get an ideal diary from a great rater, that can be used in a modeling sort of way to show it to a bad rater, and say, "Now this is the kind of diary you should keep. This should be the basis of the summary."

Dr. Cascio: Sure, when we're talking about modeling, it doesn't have to be a videotaped thing; it doesn't have to be the traditional approach to modeling that we look at.

I think what I'm talking about is this--I think if this is the job performance domain, 100% job performance, and suppose we were in a prediction situation, a job selection situation, we would try to give tests and interviews, and any other selection devices that would tap unique portions of the variance. Granted there would be considerable overlap between these. But our basic idea is that the more relevant job performance variance that we can tap, the more criterion variance that we can tap, the more valid our predictions are going to be. And the same approach holds true with performance appraisal.

So in a sense you can talk about the diary keeping or the observations, the critical incidents approach, as simply methods of gathering more and more job relevant information on which to base a decision. To go back to the LIAMA interview study, those people who didn't write down anything that was happening in the videotape essentially had probably just a little, very circumscribed idea of what actually took place. But those people who did take extensive notes and watched exactly what happened had a better sample of information on which to make judgments.

Dr. Kavanagh: I think there are certain types of things for which you can use the Goldstein and Sorcher type of program. The program is designed to model behaviors, not to model processes. So what Wally is doing in his study is looking at processes and I think that's what his question was. It would be difficult to model that. But if you use the same basic four-stage process as Goldstein and Sorcher, it doesn't matter. You simply develop tapes or movies to improve people's observational faculties. You keep repeating similar stimuli, and you make them more and more complex until the trainees are proficient, and then you reinforce them. They role play in front of the people and then you reinforce that behavior. But so much of that implies that we know what it is that makes a good observer, and I'm not sure we know that from the person perception literature.

Dr. Cascio: We do know that the more accurate information the observer collects the more valid decisions he's going to make.

Dr. Kavanagh: I'm agreeing with that. I agree with recording information, over time.

Dr. Bernardin: I agree with that, of course, the modeling. But not with a straight modeling approach where people view videotapes of other managers rating behavior and then the participants are supposed to model the behavior of those managers. That's the approach that I think is just too contrived to work because they really don't get into the rating process at all. They're just watching people rate lower or with more variability across dimensions, that sort of thing.

Dr. Cascio: Again to summarize what we said. I think the reliability of performance doesn't even become an issue until the reliability of judging performance is adequate, and that's what we have typically neglected to date. Ultimately, where are we going? Our ultimate aim is to develop decision systems; so measurement and prediction are only steps on the way to the goal, and the goal is to make decisions about people. Ultimately I think we have to ask ourselves, whatever system we develop, whether we are paying less human and social and organizational costs using that system than we are at present using whatever system we're doing now. If the answer to that is, "No, we're not paying as much cost; there is a payoff from it," then use it, because our ultimate objective is to make decisions about people. I think we need to see that; otherwise, we can't see the forest for the trees if we put ultimate stress on measurement and on prediction rather than on looking at where we are going with those. Well, we're using those to help make decisions about people and that's the bottom line.

Dr. Mullins: A point I wanted to make was that I was fascinated with your fascination with acceptability and this has come up with us a number of times too. As a matter of fact, I recall a minor explosion in our group some months back when we were first really wrestling hard with this problem, and I proposed only half facetiously that the only way to do that is just let the people get together, let representatives of the union, and the workers, and the management get together, and tell us, "All right, this is what we're going to evaluate on." And if it's standing on your head, we don't care. What do we care? As long as it satisfies everybody concerned, that's going to be it. We modified that position somewhat a little later. At any rate, that did seem like a possible solution.

Dr. Cascio: I might also add that on that performance testing experiment, we were very open about what it is that ompetence means. We found results that ran very contrary to a lot of those that are typically reported in the literature about the biases on sex and race and age. All of those things seemed to wash out when the standard was very explicit.

Dr. Ree: A couple of months back there was an article that appeared--I can't remember where it was; I read it and passsed it on to Dell Toedt--dealing with acceptability of performance rating systems, or with rating systems in particular, in a policy capturing approach. Do you remember who the author of that was, Dell? It seems to me that one of the things that we could possibly do, and it might be done in any rating system, is to look for the factors that do shape people's judgments toward the system and to promote those factors. I'm not saying we should propagandize them, but rather we should lean very heavily on those things that create a positive influence.

I think that in Government we have more constraints than virtually anybody else. Not only do we have all the EEO type constraints that one sees out even in large companies, but we're constantly under scrutiny. We have to be that much better; we have to be purer than Caesar's wife, and I think that by the proper use of these facets that have been identified as shaping attitudes toward rating systems, we ought to capitalize on them.

Dr. Kavanagh: I don't think you should apologize for propagandizing. Any good OD (organization development) effort does that. The Army calls it OE (Organizational Effectiveness).

Dr. Ree: May I make one comment also about the McClelland article? I use this one in one of the courses that I'm teaching because I think it's an example of a slightly misleading article. I don't mean to be nit-picking, but I think the article is slightly misleading in some ways. I think it's good for us to remember that when we want to talk about secrecy, there is at least some reason for secrecy in terms, say, of testing keys. If we want to have entry level tests for enlistment, etc., that's a very vital issue. McClelland takes a pot shot at tests, for what reason I'm not quite sure yet. But he starts talking about grades and testing and how the people that got C's in the college he went to only got into less well known medical schools. He was teaching at that time at a university, an extremely selective school, so that the variability of the individuals there was probably extremely small. But I think we want to be careful with that.

One of the things that compelled us in the early 70's toward work sampling, which I think must be like performance testing, was that it had a lot of face validity and it seemed that if someone could do the thing that you're asking them to do, it didn't matter whether they learned it for the test or knew it for 20 years; what you wanted was a demonstration of ability. And in that case, you absolutely and completely have openness. You say to the person, you gotta make this widget. I think those distinctions must be kept in mind.

Dr. Cascio: We use performance testing for some non-traditional types. Traditionally, you use it for plumbers or carpenters or electricians or people like that. We also use it for planners and for accountants and for management type jobs.

Dr. Ree: We had extended at that time into machinists, into secretaries, into bank tellers. Bank tellers made a very interesting test.

Dr. Cascio: I like your idea about the necessity for not compromising test items. I believe that 100%. And in these fire fighter exams that I'm developing I'm spending a tremendous amount of time getting acceptability of the system for the people who are going to be affected by these. We're spending time using three different job analysis methods to get them to tell us what they do, what is important on the jobs, using task analysis, functional job analysis, and interviewing. And then we'll go back to them and say, "Well, this is what we got. Does it correspond to what you're doing?" Ultimately we'll have the best picture that we can get of exactly what they do, getting them involved, getting their participation. Then we want to develop a reading list, a set of materials, a content sample, if you will, of what they consider to be job relevant reference materials and from those, test materials are developed. Anything that's drawn from these is fair game. Now, the test questions themselves are never disclosed, but they know where the test questions are coming from. They have ample access to see the job analysis data; they know how we're going to weight the various portions of the test because we tell them. Those tasks that seem to be a lot more important are going to get a lot more emphasis on the exam and you can expect to find them there. I believe that that kind of an approach, with more openness, is going to pay off in the long run.

Dr. Kavanagh: This whole idea of work sampling and performance testing brings us back to the Campbell et al. assessment center approach, which is what you talked about in your paper. They did a good job of assessing performance and potential. The question is whether we have the kind of money to do that. Wally, I think you said that the rating of overall potential was the best predictor. That's because of the kind of analysis that they did. It's the most efficient predictor but it wouldn't help you understand what happened in terms of managerial development.

Dr. Bernardin: One high ranking civil rights official, by the way, on that issue at a Guidelines conference last December said ratings of potential are out, we can't use them, there's no way they'd survive in court. The only way you could do it, I imagine, is if you had something like a job analysis over levels, and then identified common elements. This seems to be the language we're supposed to adopt now, since that's what the Civil Service Reform Act is using, job elements. If you have common job elements, then you can rate on those dimensions. Don't rate for potential, rate for past, but then you can argue that that's a predictor, but you don't rate on potential.

Lt Col Ratliff: Rate the past?

Dr. Bernardin: You're rating on past performance on dimensions that overlap with the next level. Then you'll survive.

Dr. Cascio: You can still make those ratings of potential but they don't have to be disclosed to employees. There's been a court ruling on it. I just read that recently. It was a private sector corporation. Oh, I know where that was; it's part of the privacy act. There was an article in the Personnel Administrator last year on the implications of the Privacy Act for personnel management.

Dr. Kavanagh: I want to emphasize a few things that I think Wayne did not emphasize. First of all, he talks in the paper about the need for understanding of performance. Even though there have been some advances made, I just don't think we understand it. That struck me as one of the major points that Wayne was making here, that we don't understand, and what we don't understand we can't measure very well. The reason is that we haven't looked at things like variability in time-series designs in the measurement area.

I liked some of his ideas on criteria particularly the notion of social utility. We try to use something which we can accept and use as a performance measurement system that is better than any other alternative, and I think that's the kind of judgments that we can make. In personnel work, we are looking for something that is better than, more cost effective than, the alternatives.

Looking at variability as a criterion, I do not feel that what you called intrinsic variability or intrinsic sources of unreliability are nonpredictable. Our research paradigms have been crummy up till now. What we need is to be getting into some of the things that the people in chemistry and finance are doing with autoregressive functions. The formulas are there but the blasted computer programs are not yet up to snuff. This is some of the work that Box and Jenkins have done. We can probably look at what we classically have defined as unreliability and be able to measure it in a systematic way with good time-series data. But that is a way off. I don't think you're going to be interested in doing that with your present system, anyway.

The notion of decision emphasis, that we are looking at a decision system, and that has been my emphasis most of the morning, is right on target. I mean we're looking at a decision and that ties in with your idea of openness, secrecy, etc., and with some of the ideas in my paper.

I happen to think that as we move towards an ideal state in an organization, there is no appraisal system because the people themselves know whether they're doing a good job or not. I'm particularly intrigued with your bus driver example with 100 points and working their way down, because that meets all the criteria of shortening the time between performance and feedback and makes the feedback clear and objective. I'm not saying this is a perfect system. I'm simply not surprised that the people endorsed it and that people know that they are self-regulating their performance. I think that there is literature in clinical psychology indicating that people do not like authority figures telling them how well they've performed.

Goldstein and Sorcher's ideas that you addressed also come from the same basis. If you talk to Mel Sorcher, he talks about this four-step process but he says, the most important thing in supervisory behavior is that we maintain the self-esteem of the subordinate. I think that's critical, something that we often overlook.

Dr. Mullins: Before you start your paper, in response to one of your comments on the previous paper, I think we ought to keep distinct in our minds the difference between an operational system and a research system. We have to do that here all the time. Saying that probably cost effectiveness evaluation of a system is more important than pure evaluation such as Wally is doing is getting into some of that confusion. I think what Wally is looking at is an understanding and that, I think, has very little to do with cost effectiveness. If we're talking about an operational system, then cost effectiveness is extremely important and the pure research then goes on in another world.

Dr. Kavanagh: That is why I continue to do pure research as well as applied research. I also see an interrelationship between the two which is what I address in my paper.

CHAPTER 5

PERFORMANCE ASSESSMENT IN ORGANIZATIONS: SOME NON-RANDOM OBSERVATIONS

Michael J. Kavanagh
School of Management
State University of New York at Binghamton

Prepared for AFHRL Conference, March 19-21, 1979, San Antonio, Texas.

I. Performance Feedback

While introspecting on the request from Dr. Mullins to present a theoretical paper that is a "broad, speculative, and personal position paper on the state of the art of human assessment," it struck me that the central concern of a performance evaluation system in an organization must be the feedback function involving individual employee performance. As will be discussed later, it is argued that all other parts of the system (e.g., the type of format used, the stimuli evaluated, the purpose(s) of the system) either directly or indirectly impact on the quantity and quality of the feedback employees receive. But first, let me share several thoughts that led me to emphasize the importance of this function over other parts of the process of human assessment within organizations.

First, current prescriptive techniques (organizational interventions to improve employee motivation, quality of work life, or both, all appear to include an improvement (often, a reduction in noise) in either the quantity or quality of the feedback the individual employee can obtain concerning his or her performance. In most cases, this is achieved by making the feedback loop shortened in time and by improving the visibility or objectivity, of the performance feedback information in the organization.

Several examples are readily apparent. The implementation of a behavioral modification plan (see Hamner, 1975) in an air freight concern involved, as a critical element, that employees keep records of their own performance for comparison against established standards or goals. In an automobile manufacturer experiment (Editor, Organizational Dynamics, 1973) involving the use of autonomous work groups, the quality of feedback was enhanced, in that the workers could see the finished product rather than a single piece. A key element in job enrichment programs (Lawler, 1969) is that employees exercise greater control over more elements of their individual jobs. This would usually improve both the quantity and clarity of feedback. Management by Objectives (MBO) and other goal-setting programs have, as part of their design, the identification and development of organizational data against which performance goals can be evaluated. In general, it would seem that the desired end state of these motivational programs, if one extends the logic involved, would be self-regulation of individual performance through improvements in the quantity, quality, and clarity of performance feedback data.

Another trend in current years that emphasizes the role of the feedback function has been the emergence of Behavioral Anchored Rating Scales (BARS) to improve the assessment of human performance. Although the evidence on their assumed psychometric superiority has been mixed, and serious questions remain regarding their cost effectiveness, one thing seems clear: employees (rater and ratees) like BARS more and are more likely to use them as compared to traditional trait ratings. One can't help but wonder if this increased attractiveness is not a function of the fact that BARS performance dimensions provide clear (unambiguous) feedback on employee performance.

This linkage between the quality of the development of a performance evaluation system and the quality of the feedback will be more fully seen in terms of a sequential model. Another reason for emphasizing the importance of the feedback function is theory-based. Following Maslow (1965), among others, I would argue that people want to effectively utilize their personal abilities in fulfilling their job requirements. However, as Festinger (1954) has further noted, not only do people want to use and evaluate their abilities, they also want to find that they are good. This "goodness" in their use of their abilities in their jobs can only occur via the feedback function. As a footnote, although the previous statement appears to be almost a "truism" in terms of individual psychology, it is also a "truism" that not all employees can become the Chief Executive Officer, and that most appraisal systems in organizations emphasize "improvements needed" or force the feedback of negative information under the guise of "growth and development."

As a final thought on the importance of performance feedback, I will draw on personal, anecdotal evidence. In various relationships with organizations either for research or consulting purposes, I make it a point to interact with employees at various levels in the organization. Two questions that I find most revealing involve performance feedback. These are: "How do you know you are doing a good job?" and "How do you know when you are doing a bad job?" Most people cannot answer the first question--they just don't get that kind of feedback. On the other hand, most can answer the second question, and the most typical answer is "When my boss yells at me." Such is the nature of feedback in organizations.

Now I will turn to a description of a sequential model that involves the important linkages in a performance assessment system. The model includes two categories of characteristics, direct and indirect, so named because of their effect on the type, quality, and quantity of feedback. The aspects that would have direct effects on individual performance feedback are: (1) rater training; (2) a goal-setting or MBO system; (3) whether the performance evaluation data are based on objective and/or subjective standards; (4) for rating scale data, the descriptive clarity (e.g., BARS) versus purely numerical data (e.g., simple trait rating graphic scales); (5) the degree of correspondence between the performance system and the reward system in the organization; (6) for ratings, the source of the feedback data used, i.e., superior, peer, self, subordinates; and (7) the comparison standard for individual data, i.e., normative versus ipsative.

Somewhat more indirect in terms of their impact on performance feedback are: (1) the traditional psychometric properties of reliability, validity, and freedom from bias; (2) the "practicality" standard, but only in terms of time requirements to evaluate employee performance and the time requirement for the feedback interview; (3) the major purpose of the performance assessment--administrative versus employee growth; (4) the general managerial philosophy in the organization; (5) the presence of a union; and (6) the organization's sensitivity to EEO concerns.

The reader should realize that the preceding lists do not exhaust all the factors that will impact on the performance feedback function, but they certainly represent some major ones. Furthermore, in order to place these aspects of the organization as they impact the performance assessment system in perspective, one should envision a sequential flow diagram with the indirect effect factors flowing into the more direct factors. These, in turn, affect the quantity, quality, and type of individual job performance feedback. Continuing this model, there would be a direct effect on changes in job behaviors from the feedback, and, in turn, effects on job attitudes flowing from the changes in job behaviors.

It should be apparent from the flow model that when there are incongruities in the system, the quality of feedback will suffer and this will affect the remainder of the model. For example, an organization with an "autocratic" managerial philosophy that would emphasize employee growth through the use of ipsative performance profiles and goal-setting interviews will probably cause a personal conflict for supervisors, who then will be forced to "beat the system." Further, this discussion and model indicate that a given managerial philosophy and stated purpose for the performance assessment system will greatly guide the development of the remainder of the system. As a final point, it should be apparent that in designing a new performance assessment system (or revising a current one), the initial step should be the diagnosis of those indirect aspects that will impact on the feedback function.

II. The Performance Appraisal Interview

After considering the role of feedback in the performance assessment system and the development of the sequential model described previously, my introspections turned to the delivery of the feedback. Prior to this discussion, and lest the reader misunderstand the thrust of this paper, I would like to briefly address my perception of the state of the art of human assessment. As the feedback model implies, the traditional concerns of measurement, rater training, and psychometric evaluation of the performance measurement have not lost their importance. They are still quite important, and, in fact, by tying them into a model of feedback, their importance has been increased. That is, I am not advocating that we abandon our concern with these more traditional topics, but rather, they must be put into a systems concept based on the importance of performance feedback to the individual employee.

Regardless of whether the performance appraisal interview is used for administrative or employee growth purposes, the same problem of communication between supervisor and subordinate exists. It seems obvious that the focus of any efforts to improve the interview should be concerned with the communication of negative information about performance. Communications about positive strengths of the individual are very easy to do, but generally are not helpful in improving performance. The difficult part of the appraisal interview is how to give the "bad news" about employee performance. I am not suggesting that positive performance information should not be communicated since good performance must be maintained; however, various attempts suggested in the literature to do both within the same interview have been failures. Thus, we need to take a close look at how the appraisal interview should be conducted in order that feedback is given that will aid employee performance.

There are several procedures that should be useful in improving the performance feedback interview, particularly in terms of giving negative feedback:

1. Training programs for supervisors that focus on emphasizing the threatening nature of the performance interview and the basic human need for self-esteem. The training program should be experiential, involving role playing of real situations and, hopefully, using videotape recording so that trainees can view their practice at interviewing. A training program modeled after the behavioral modeling approach (Goldstein & Sorcher, 1974) with its emphasis on maintaining self-esteem of employees would be most appropriate.

2. Institute a formal procedure in the organization whereby employees rate their own performance as well as having the supervisor do the rating. Then, the employee and supervisor would exchange their ratings prior to the appraisal interview. There is full sharing of information, and the differences in perceptions relative to the performance of the employee are known before the interview begins. As an added twist to this approach, the employee and supervisor would complete two judgments (in pencil). The first would be the typical evaluation indicating the ratee's performance level on the performance dimension, trait, or behavior. The second rating would be a certainty evaluation, indicating how certain the rater is that his or her first rating accurately describes the performance of the individual. This procedure should lead to interviews with much higher task-relevant behaviors and reduced emotional behaviors, particularly in terms of negative feedback. It should be clear, however, that these procedures are based on the assumption that the individuals will be honest in their ratings.

3. The notion of honesty in ratings, I believe, is the critical variable to make the appraisal interview an effective management tool for administrative decisions and employee growth. Here the conflicts between organizational reality and the "truisms"

regarding human performance in organizations, alluded to in Section I, are very relevant. We cannot expect honesty in the supervisor-subordinate dyad if honesty and trust do not exist throughout the organization. Even if the interview is concerned only with employee growth, and the completed forms are kept in the supervisor's desk (not sent to Personnel), we cannot expect honesty to exist if the organizational climate is such that it does not support tnis.

4. In order to improve the performance appraisal interview, particularly as it relates to giving bad news, suggestions concerned with the redesign of jobs seem most appropriate. That is, building better feedback mechanisms on individual performance within the job will ease the burden of the supervisor "surprising" the employee at appraisal time. Clearly define the organizational data that will be used to judge employee performance. If none exists, then create it. If the data are objective, start recording them, and have the record available to the individual employee. If they are subjective (supervisor's judgment), be open and tell the employee this fact. Of course, the difficulty with this approach is that redesigning jobs takes considerable time, particularly managerial ones.

III. Summary and Escape Valve

I have discussed a number of ideas in this paper. My main concern is with improving the performance feedback data employees receive. This should be the major goal in the design or redesign of any performance assessment system. I am convinced that feedback on performance, whether given on an event-by-event basis or once a year, does impact on employee behavior. Too often this impact is negligible or negative--we need to improve our performance assessment systems to increase the positive aspects of feedback.

Finally, my escape valve is that I am dating this paper March 1, 1979. This means I am free to change, modify, strengthen, etc., any ideas presented here anytime in the future.

References

Editor, Organizational Dynamics. Job redesign on the assembly line: farewell to blue-collar blues. Organizational Dynamics, 1973, 2(2), 51-67.

Festinger, L. A theory of social comparison processes. Human Relations, 1954, 7, 117-140.

Goldstein, A.P., & Sorcher, M. Changing supervisor behavior. New York: Pergamon, 1974.

Hamner, W.C. Reinforcement theory and contingency management in organizational settings. In R.M. Steers and L.W. Porter (Eds.), Motivation and work behavior. New York: McGraw-Hill, 1975, 477-503.

Lawler, E.E. Job design and employee motivation., Personnel Psychology, 1969, 22, 426-435.

Maslow, A.H. Eupsychian management. Homewood, IL: Dorsey Press, 1965.

Dr. Kavanagh: If I can get a system that has high relevance and I can convince my raters that it has high relevance--that it's accurate and valid--I would increase their confidence in the system and increase its acceptability. In turn, this would mean increasing the use and increasing the cost effectiveness in a practical sense. That's the tie-in I see.

There are two separate sets of standards that one uses. The first thing you must do is determine what the political climate is in the organization and what the organizational philosophy is before you can even start to design the new system. I have for years been wringing the variance towel, trying to get the greatest amount of variance out of the performance measures. I have been moving away from that, trying to move towards how we deal with performance assessment systems.

What are the important issues? I'm not belittling pure research. I simply think that we have a lot of people working in basic research, but we have very few good scientists working on the implementation side. Most of that is very soft; most of the measurement is very weak. So when my rubber was hitting the road, I was introspecting, as I point out in my paper, and I tried to develop a model. I said to myself, "What is important to me, what kind of feedback do I get on my job? How do I relate to performance assessment systems?" I realized that the main way I relate is if my supervisor comes to me and tells me something. So I decided to focus on feedback, and start thinking about the feedback between supervisor and subordinate. There's a lot of good literature that indicates that most of the new motivational techniques that we have, and certainly a large number of the new faddish managerial training programs, all have one element throughout them. They either make the feedback to the individual about performance cleaner, clearer, or faster. That's a common element in all of them.

So, where does that leave me? Well, it says that we've got to clean up our feedback systems and the way that we clean them up is to posit a model as I have in my paper. It is a sequential model. There are some indirect effects on the quality of feedback (my dependent variable), a feedback that the supervisor must give to the employees. And there are some indirect effects.

I talk about organizational philosophy; I talk about the presence of a union, and the need to comply with EEOC. There are six or seven indirect variables, and these, in a sense, flow into more direct kinds of things such as rater training. Rater training, as opposed to no rater training, will have a direct effect on quality of feedback. The presence of an MBO system, or objectives, or a goal setting system will have an effect on the quality of the feedback the individual receives as opposed to having no goal setting system.

Dr. Cascio: Objective vs. subjective data?

Dr. Kavanagh: Objective vs. subjective data would be another such factor as well as sharing vs. non-sharing. Sharing the performance appraisal before the interview vs. non-sharing I think has an impact.

The quality of feedback is directly related to behavior change. If the quality of feedback is poor, you're getting no behavior change, or negative changes.

This assumes you want behavior change as a goal. You might want to just maintain. I recognize that in some cases, the worker is performing at a relatively high level, and you want to just keep him at that level. The model will handle that situation also.

Then finally, there's change in job attitudes as a result of behavior change. I put it that way because I think that reflects the literature. I'm also working with a mediating variable called "confidence in the system." I assume that if you use something like Bales' Interaction Process Analysis, you can talk about negative emotional behaviors, positive emotional behaviors, and task relevant behaviors. The ideal feedback interview is one in which there is a great deal of task relevant behaviors and low amounts of positive emotional and negative emotional. That approach will increase the ratee's confidence in the system of performance appraisal. Now this flies in the face of a sandwich approach to performance feedback. You give some good, you give some bad, and then you give some more good. What I'm saying is it should be task relevant primarily.

If you consider this approach, and want to increase quality of feedback in those three ways, how do we do it? I looked at two things, the sharing and non-sharing of performance appraisals and the level of previous performance of the individual. The sharing and non-sharing definitely impacted as a main effect. We found that by sharing ahead of time, we were able to reduce emotional behavior and increase task relevant behaviors in the interview. We did this in a video taping situation where we taped the interviews, then had raters rate on Bales' categories. Interrater reliabilities were in the .9 range. However, that research is just the first step.

One of the things that bothers me is that when I talk to people in organizations about the performance feedback that they get, they tell me they get lots of negative feedback. They know when they're doing a bad job but they don't know when they're doing a good job. It means that we're doing a terrible job in terms of positive feedback on performance.

In my paper, I point out the real problem is that we need to train people in interviewing. I'm not talking about training raters. I am talking about training interviewers, and there you can use the Goldstein and Sorcher approach in training people how to give negative feedback, with an emphasis on maintaining self-esteem. Students in my interpersonal skill class have been doing this as class projects for the last 2 years. Some are failures, but others handle it well.

There are several other ideas I like. I like sharing the results before the interview. I like people rating their own performance in pencil. I think people should rate twice. First, in terms of where they think the person is on a particular dimension, and then they should rate in terms of their confidence in the rating. The ratee should do this, also. I think that provides a good basis for discussion in the interview.

One of the things that I've been doing with a company for which I am developing a performance appraisal system is to make the performance feedback loop shorter, purer, and cleaner for the employees all the way from production line to executives. I've been doing that using a modification of the work standards notion, modeled a bit after the Department of Labor performance evaluation system. The first column of the rating form has the duty requirements very similar to the Department of Labor system. The next column has results expected and the results expected are specified in terms of levels. The third column, which is a little different than most people have, has organizational data available against which to judge the performance. In a lot of cases, organizational data have to be created. This is not a solution to all problems in appraisal, but this kind of organization has helped. Furthermore, employees know how well they're doing. There are records there.

Lt Col Ratliff: When you say you're using the Labor Department's method, are you talking about their old method before they junked it?

Dr. Kavanagh: Well they've just modified it a bit, haven't they, as I understand it?

Lt Col Ratliff: Well, when I was up there in June what had happened was that they had turned over to their Union, or had agreed with the Union, that they would only use appraisal from that method in consideration for promotion. It really escalated ratings through the top and so they said they were junking it and developing a different system.

Dr. Kavanagh: I'm extremely familiar with the old system. The new system I'm not totally familiar with yet, but it still has duties and end results expected in the first two columns. That's very common to most MBO systems. I think the third column is a little different. I feel strongly that the supervisor and the employee should develop organizational data together, agree upon levels of performance, and what the data will be that will be used to judge what the level of performance is. I don't know how accurate that is going to be, but it has a lot of what I call "organizational" validity. The last column is comments and goal setting. Supposedly, there's room for improvement in all cases.

In the organization that I'm working with, the form never goes forward to Personnel. It stays in the supervisor's desk, and there's a single page that goes forward.

Straight, flat-out administrative ratings that are distorted, and oftentimes bear no relationship to the first form--that's typical of most systems I've seen. Managers have always said to me, how can I tell the truth about this individual? They won't get anywhere.

Dr. Borman: Have you done any research with, or made any observations of, confidence ratings? Are there, for instance, rater main effects where some raters are just absolutely not confident about anything? That may be correlated with how close they are in terms of the organizational level or it may be correlated with how much time they spend with the employees that they're in charge of. It seems to me that that would be really interesting data.

Dr. Kavanagh: We're planning to do something on that as soon as we can. My plan is to research it with some fairly basic things like known or at least scalable art objects--making people rate on dimensions of those kinds of stimuli, rather than people. I will get to people at some point, and then ask them about their confidence in their ratings.

I have this feeling that the psychometric work that we do is absolutely necessary and critical. We can never get away from that because we must demonstrate to management that it is a good system. Otherwise they're not going to have any confidence in it. But we need to deal with confidence, also.

Dr. Mullins: It seems to me that one thing you could do immediately with that confidence number is very simple, but I think it would be an interesting thing to do and that is to correlate that confidence number with the accuracy of the rater against some external criterion. I wonder if the more accurate ones are the more confident ones, in other words. I have no idea whether it would be or not.

Dr. Ree: There's a fairly extensive literature in confidence weighting of test batteries that I think may be relevant to this particular discussion. The Air Force in the 1950's sponsored a couple of contracts, the title of which was something like "Permissible

Probabilities Scoring." Something along those lines. We also did a number of things where we looked at confidence for weighting any number of other things and it seems to have certain effects. They seem to have one uniform effect of making scales univocal, but univocal to the point of killing important variability in the scale. That is, they get rid of unique variance that allows them to predict other criteria.

Dr. Kavanagh: I conceive of the confidence ratings as a separate dependent variable, rather than a weighting type system. Jeff Kane developed a system that is essentially a "take-off" of confidence. What I am referring to is called certainty scaling, which Lee Wolins has developed. That's really what we are saying, "How certain are you?" But rather than using it (as Lee does) to change the underlying distribution of score responses to items, I'm saying that it is a legitimate, separate, dependent variable in the system. In my model, it is a moderator variable.

Dr. Ree: I was about to say, "Wasn't work done by somebody by the name of Al?" I forget what his first name is.

Dr. Cascio: Et?

Dr. Ree: Well, in the early 70's there seemed to be a whole flurry of activities in moderated regression, moderated variable research, that looked at confidence among other things.

Dr. Kavanagh: I will find that literature. We just stumbled on this, and I've never stumbled on anything that's new.

Dr. Ree: That's all right. I don't think any of us ever do and that's not really an important fact of life. I think it covers roughly the period from about 1970 through 1974. Whoever wrote that particular article in the 1972 or 1973 Annual Review of Psychology paid a lot of attention to it.

Dr. Kavanagh: Thank you.

Dr. Muckler: I'd like to comment on four points in Mike's paper. The first is rather heavily stressed in the paper on the first page, and that concerns the motivational aspect of some of the things leading to "Self-Regulation of Individual Performance," etc. Wayne, if I might comment on something you said, you made the comment that the goal of the performance appraisal is to make decisions about people. That's certainly true, of course, but as a supervisor, a probably more important goal to me is that this is one of the tools that I have in communicating with people and hopefully I will go through my chain and hopefully I will affect behavior, and hopefully I will affect it positively. I'm not always sure that I do that, as a matter of fact. The action research people have been really beating on us lately and particularly Lou Davis with the Quality of Work Life Study. He was describing a thing he is doing right now with a plant with 2,000 workers. They have six supervisors in that whole plant. They have destroyed, they have wiped out the whole middle management structure. It's all on a model with self-contained teams. They have no performance appraisal. The feedback is internal team feedback and the results that he's obtaining are just really spectacular. He says it's really rather difficult in the general organization because this plant stands out so much with respect to the rest of the organization that now a top management decision has to be made--are they going to diffuse this to the rest of the organization? But I mean literally 2,000 people and six supervisors, and no performance appraisal system at all.

The second point I'd like to comment on is the question of the criterion evaluation. I want to follow up on what Wayne said, and that's the question of acceptability, or to put it this way, does it appear fair? And I'd like to comment not just on does it appear fair to the individual but does it appear fair to the supervisor as well? We have done some informal followups on what was said in performance appraisal interviews and the only thing that I can conclude in some cases is that it was two decorticate individuals talking to each other where stress has peaked out. And this is particularly true with younger supervisors. I really have to watch younger supervisors when they go and do this for the first time because the threat that's on them, if they're at all sensitive, gets to be really extraordinary. One young lady supervisor which I have, as she approached her first performance appraisal as a supervisor, was in tears and was going to quit. She didn't want to be a supervisor anymore. Now that was a little extreme, of course, but we also try to follow up for the younger supervisors when they do an interview what was said, and I will interrogate both the supervisor and the employee and you know what you hear. They weren't listening to each other. They're hearing different things and you know they're saying different things. I wished I had the courage to tape some of these. I couldn't do it, but I'd really like to tape interviews because what comes out of it is just bizarre.

And that's really the third point, this problem of communication between supervisor and subordinate. If we're going to maintain the self-esteem of the individual which I think is absolutely correct, I'd also like to maintain the self-esteem of the supervisors as well, because many of them do get really threatened by these particular situations. In that part, I wish I knew, Mike, how to deal with failure, because the first problem is that nobody admits that they have failed. At least nobody I've ever given an interview to. What's more I think they're fairly firmly convinced, rightly or wrongly, that they did rather well under the circumstances. So I have had a great deal of difficulty with this, and I think everybody has great difficulty with it, except for those insensitive psychopaths who can say you're a lousy guy. But there aren't many of those and I'm not sure I'd want him to be a supervisor anyway.

My last point is a question that Mike brings up on honesty in organizational climate and we've been talking about honesty and openness. To go against honesty and openness places me in a very difficult situation. I must confess I have had second thoughts about that over the years. Let me put it this way, when I was a young supervisor I felt very strongly in communicating everything. But then I discovered that a lot of the information that I had was noise, that it represented management turbulence, or rumors, and I found rightly or wrongly that I had begun to screen a lot of information that comes down the line. And I really haven't made up my mind whether this is right or wrong. I do know that if I communicate fully to them I've got an unstable, hysterical work force. It's just that a lot of it is frightening. But one question is, "Should you lie to people when they're not doing so good?" What's the best motivational strategy to deal with people like that? Is honesty the best policy? Should you say to them, "You're lousy, you're doing a rotten job?" Or is maybe some selective lying more motivational? Frankly, I haven't decided yet.

Dr. Kavanagh: Relative to the Goldstein and Sorcher study, I've seen some of Mel's tapes, and he deals with a specific problem called the poor performance interview. The approach they take stresses the maintenance of self-esteem. When you maintain self-esteem, you don't personalize the problem. That's assuming this is a performance problem that has occurred with an employee who has previously been performing at a fairly good level. The boss sits down with him and says, "Now we have this problem." He emphasizes that if a subordinate employee is performing poorly, then it is a shared responsibility. Sorcher has had some phenomenal success with that type of training, with role modeling. I was impressed by the tapes, and then to see some of the supervisors who cannot do that. It's not in their behavioral repertoire to do that. They just can't sit down and say to an employee who's working for them, "Your poor performance is part of my problem. It's my problem because I'm not supervising you properly."

Dr. Cascio: I'm not sure I have the answers on what to do about somebody who's doing poorly but I think I have the answer about what not to do. And that is save it up. I don't think there's any more emotional performance appraisal than when a student comes in at the end of the quarter, and they've got to pass your course to graduate, and you have to tell them that they did fail, and tell them why, and have them leave the office not feeling good about it. What I do as a result of getting burned on that a couple of times is that when I see ineffective behavior in the classroom or after an exam, I don't save it up. I tell them right on the spot. I take them aside and I encourage them to come and see me and I tell them immediately. I believe in this immediate feedback bit because at a later date we want these people to believe that we're not attacking them as a person but rather we're attacking their behavior. And so it should be immediate, not saved up.

Dr. Muckler: Frankly I've always found that unconvincing.

Lt Col Ratliff: I have one question on openness. If you're very open and you lay out a set of behaviors or performance standards, people have different ways of interpreting these behaviors. When you get down to the rating time and there is a problem, they may say, "It's what we agreed upon." You reply, "Yes, but at the time it had a more specific meaning and things don't fit our agreement very well." What do you do in that case depending upon where people are coming from, defensively, and you find some of the original standard not met? I don't think communication is always quite that complete. I think openness is good, I generally support it, but it does have a certain number of pitfalls, if openness is designated at 100% rather than a lower level. If you've got, say, 85%, then you can change with circumstances.

Dr. Kavanagh: You're saying if the supervisor is totally open, and if supervisor and employee set mutually agreed upon standards of performance, and then the employee says I met these, here's the objective evidence. However, if the supervisor at that point says, "Yes, but, . . . ," then that's a pretty poor supervisor.

Lt Col Ratliff: No, what I'm saying is that the employee did not necessarily meet what you thought was a clear understanding of standards or behaviors. You know--if you say we want a technical paper and he brings in something pretty sloppy, you know your standards for a technical paper . . . .

Dr. Kavanagh: I mean that openness should just be a greater sharing of organizational data by all members in the organization rather than I'm going to tell you all about my sex life. That's not open, in the way I mean it.

Dr. Mullins: Fred, I think it's time for your paper now.

CHAPTER 6

PUBLIC LAW 95-454 AND PERFORMANCE APPRAISAL

Frederick A. Muckler
Navy Personnel Research and Development Center

Public Law 95-454

The passage of Public Law 95-454 ("Civil Service Reform Act of 1978") on 13 October 1978 has created a number of urgent issues to be resolved. Aside from reorganizations of the Civil Service Commission and restructuring of the executive GS levels, the central issue in PL 95-454 is the recognition and reward (or punishment) of personnel performance. For psychologists, this is another event in the continuing saga of personnel performance appraisal. And psychologists should be expected to have a great deal to say about any brand of performance appraisal--some of which may even be of practical value.

Some general thoughts about this problem:

1. It is a great deal easier to say what is wrong with current practice than what might be right. "Standard practice" in performance appraisal tends to be so invalid, unreliable, and unfair that most applications are easy targets for criticism.

2. The psychological literature on performance appraisal is now about 75 years old. It contains hundreds (if not thousands) of published papers on human performance appraisal. It seems (to me) to contain a considerable store of good knowledge about what to do, and what not to do, in a real-world performance appraisal system.

3. But, as yet, it does not appear (to me) that psychological knowledge has had a significant, continuous, and major impact on practice. Perhaps it is too soon; scientific knowledge traditionally is absorbed into practice very slowly. After all, current practice is the accumulated result of millenia of chaotic personnel performance measurement. And, like all psychological phenomena, human performance measurement is bent by underlying and often ambiguous perceptions of what human behavior should be.

What is This Thing Called "Merit"

The basic assumption of PL 95-454 is that personnel actions (rewards, promotions, assignments, dismissals, etc.) should be based on merit rather than other possible criteria such as longevity on the job, height and weight, or whatever. This rule of merit is assumed to be good, a basic perception apparently shared by many (but certainly not all) in Western technological societies. "Merit," therefore, is good; it is an undefined axiom in this calculus of performance appraisal. Our job, it would appear, is to give some operational meaning to the term.

It may well be that everyone understands the concept of "merit" except me. But a recurrent question I have is: What is this thing called "merit"? Let us assume for the moment that "merit" in this context is doing something of value during a period of performance.

One question that immediately arises is: Of value to whom? The action may be of value to the employee, his or her peers, supervisors, organization, country, and/or society. Which of these are allowable to count as merit behavior by the individual?

It would appear, at least implicitly, that primary emphasis is on value to the organization and, perhaps, in a more general sense, the country. Whatever, it would seem desirable to be as explicit as possible as to what counts in the behavior of the employee.

A second question: Values perceived by whom? As perceived by the employee, peers, supervisor, management? One, some, or all of these? The psychological literature seems rather clear to me on one aspect of this question. Namely, the closer the evaluator is to the actual working situation, the more valid and reliable the performance appraisal judgments will be. This does not necessarily mean physical proximity; rather, it requires that the evaluator have detailed knowledge of what the employee has done and the conditions under which he or she has done it.

Merit Behavior Versus Outcome

One way of assessing the merit of individual behavior is in terms of the outcomes of that behavior. In short, did the behavior of the individual result in a successful outcome? Or, it isn't how you play the game, it's whether you win or lose.

The question here is: Is it to be recognized for behavior for which there is no apparent successful or even poor outcome during the period of performance? A case history in point. A division of a large company had lost 40 million dollars in the preceding fiscal year. The Board of Directors might have perceived that as an unsuccessful outcome. Instead, they voted a letter of congratulation and gratitude to the Division Manager. Why? The Division Manager had been brought in hastily to replace his predecessor when cost projections suggested that the division losses might be sufficient to bankrupt the entire corporation. Among other things, the new Division Manager minimized losses and perhaps saved the corporation. At one level of measurement he was a failure ($40 million in losses); at another level of perceived merit he was an outstanding success (saved the company).

A significant percentage of the civil service manpower pool is people who provide services. It is often difficult to measure the outcomes of those services. This is particularly true of research and development specialists, a major category of service personnel in the Federal government. The gap between R&D and operations can be a long one. The history of science is filled with examples of long delay between discovery and application. Mendel's work on plant genetics is the most often cited case; over 40 years between publication and recognition. Doppler's original paper was published in 1926; it was not until over 10 years later that the practical implications of the Doppler Effect were perceived. Even ir the most applied work, successful outcomes may be years in the future. The Titan and Minuteman ICBM systems have been operational for over a decade. Yet, no one knows if they will be operationally successful--simply because the opportunity to use them has not occurred.

There are some unsatisfactory alternatives here. First, we could delay rewards to R&D people in the Department of Defense until the outcomes of their work were operationally demonstrated in combat. If so, there may be no rewards for R&D people in DOD ever. (An alternative which might be quite satisfactory to some operational people.) Second, we can reward R&D people only for outcomes that have immediate acceptance and application for a present problem. By "immediate" is meant at most a year or within the remaining term of the current office holder (in the Navy, "on my watch"). The R&D community is under constant pressure for this kind of immediate outcome. There are, however, some (deluded, no doubt) who believe that the R&D of today is for the systems of tomorrow. These same fools think that concentration of R&D resources on today's problems may even mean no systems for tomorrow.

Whatever the case may be, it would appear to be desirable to be as clear as possible to the employee what outcomes are valued by the system and when. If the standard is immediate, perceptible, and acceptable outcomes relevant to today's problems, that ought to be clearly stated--in writing.

To me, the greatest single cause of trouble in a performance appraisal system is ambiguity. The more ambiguous the system, the more confused (and probably irritated) the employee becomes, and the more difficult it becomes for the supervisor doing the evaluation. The fundamental requirement of every performance appraisal system is that it state clearly and specifically what is expected of the employee. This is particularly true of (1) dimensions of desired behavior and (2) standards of performance for those dimensions.

Criterion Deficiency: Desired Behavior Sets

What do we want people to do? In my opinion we should be as clear as possible and as complete as possible in defining the categories of acceptable behavior. Yet, most commonly in practice the criterion set is both unclear and incomplete--hence my use of the term, "criterion deficiency." In 30 years as an employee in government, industry, and university organizations, only once have I seen this problem intelligently met.

Oddly enough, this was in a department of psychology in a university. The management decided they wanted to measure the psychology faculty on five dimensions:

1. Teaching effectiveness

2. Service to department, college, and university

3. Scientific and professional contributions

4. Service to the community

5. Service to the nation

A number of important qualifications were made: (1) the list is not in order of importance, (2) the list is complete from the point of view of management, (3) the individual faculty member did not have to excel in all five categories, (4) alternative ways of measuring each dimension were recommended with the ratee being encouraged to suggest other ways but in any case the particular method was the choice of the one rated, and (5) the employee was required to prepare written documentation to be used in a meeting with the rating group.

A key issue here is that the management stated clearly what they expected of an employee in terms of general classes of activities. They defined what behaviors from a faculty member were of value ("merit") to the university. In my opinion, the first step in any performance appraisal system is for the management to state as clearly as possible what it is they want. That is, to me, a necessary condition and will be the major determinant in the validity, reliability, and perceived fairness of the performance appraisal system that evolves.

There are at least three problems with this condition:

1. Management people may not know what they want and may find it extremely painful even to discuss the issue.

2. When people find out what the management wants, the people may not like it.

3. Some managers feel they lose flexibility by stating objectives in advance: "If I tell them what I want, then they will only do those things." This is management and appraisal by reaction: "I'll see what they do and then tell them whether I like it or not." This author has difficulty commenting on this management approach because he is uncomfortable with it--although he has seen it work successfully in at least one case. This was in a basic research environment which might be the only place this management technique is appropriate.

Setting Standards

Having defined what categories one wants to measure in performance appraisal, the next natural question is: How much? This is only one part of the general problem of setting performance standards, but it probably is the part that receives the most attention. In a "merit" system particularly, setting performance standards would certainly seem to be essential to help in the very practical decision of unacceptable, acceptable, and above-acceptable employee performance.

This is obviously related to the basic issue of personnel productivity, defined most commonly as amount of work produced per hour. We are often asked to believe that the fate of the nation, if not all of Western civilization, rests on ever-increasing personnel productivity. At any rate, it is doubtful that any practical performance appraisal system today would be acceptable to management if personnel productivity is not taken into account.

But simplistic views of personnel productivity must be qualified. Some of the essential hedges are:

1. Productivity, as normally defined, is an input and not an output measure from the organizational point of view. More is not necessarily better, and the question is: What is the result in product/service outcomes from given levels of productivity? There have been many examples of increased labor productivity resulting in more products than could be sold.

2. Productivity, defined in terms of quantity, does not take into account product/service quality. I know of no practical situation where quantity of production is a sufficient metric. At a minimum, quantity and quality of services or products are essential dimensions, and neither is easy to measure in any real world situation unless one is willing to accept the crudest level of counting.

A major issue for the implementation of PL 95-454 is how to set performance standards. One question is: Who should be involved in setting these standards? Many managers would appear to prefer some sort of non-participative system based on their own judgment. But the trend of modern management practice (and PL 95-454) implies some sort of employee participation in the process.

Certainly if a management-by-objectives method is used, joint employer-employee negotiation is normally expected to establish finite goals, specific objectives, and reasonable outputs. Practical experience in this procedure can make even the most humanitarian supervisor wish for a return to the days of authoritarian management, for the gap between what is "reasonable" to management and what is "reasonable" to the employee may be beyond the realm of even prolonged discussion.

Whatever, it would appear that the implied intent of PL 95-454 is that the employee should be involved in setting performance standards. The Act is not clear (to me), however, and I would suggest that this is an issue which should be discussed and resolved by somebody.

Determining Failure and Why?

Most supervisors know (or think they know) bad employees. The identification of "bad" employees is very often not so clear. At any rate, PL 95-454 requires identification of "unacceptable" performance and devotes much attention to adverse actions based on unacceptable performance. Supervisors in the system may well have to defend eventually their judgments of unacceptable performance in a Federal court, to say nothing of the intermediate hearing stages.

A line supervisor, faced with a case of unacceptable performance, might do well to ask: Why did the employee fail? In modern complex organizations, outcomes are often beyond the direct control of the employee or, for that matter, the supervisor or the agency. The employee may produce continuous high quantity and quality of outputs which have no discernible system outcomes. Is this "unacceptable performance"? Or, is this unacceptable performance not by the employee but by the system of which the employee is a part?

Many supervisors long for the days of arbitrary firing without the need for subsequent justification. Aside from the possibility that this was never good management behavior, current practice does not accept this kind of arbitrary management action. PL 95-454 certainly does not, and supervisors in the Federal government should be clearly aware of the conditions required for adverse personnel actions. In no case (removal, suspension for more than 14 days, reduction in grade, reduction in pay, furlough of 30 days or less) can the supervisor act without written justification and the high probability of appeal. An important part of the justification may be the results of past performance appraisals.

Measuring the Past and Predicting the Future

Most performance appraisal systems try to combine measurement of past performance and predictions of future performance. This should be reasonable and convenient: One may determine rewards (or punishments) for what the employee has done and how the employee should be used in the future. But this combination may be a mistake. These are two separate processes, and it may be wise to keep them apart. Optimal measurement for one may not necessarily be optimal for the other. Combining both may produce a hybrid that is good for neither.

It is a widely accepted truism that the best predictor of the future is past performance. Indeed, some would say that the past is the only valid predictor of the future. Like all generalizations, this one is subject to considerable and important qualifications. For example, the degree to which future prediction is valid depends on the similarity between past and future. The system being predicted must be reasonably stable over time. The less the system is stable, the less meaningful predictions will be. This fact is critical to initial selection for supervisory positions. Because an individual has done things well does not necessarily mean the individual will supervise well. The individual must shift from personal accomplishment to directing and planning the work of others. In short, the major components of the past and new jobs are fundamentally different.

Measurement of past job performance tends to be task-related. What did the person already do on the job? One can point to past events and specific past performance. Prediction of future performance tends to be far less clear and correspondingly far more vague--unless the specific context for future performance can be defined.

In practice, measurement dimensions for the prediction of future performance also tend to be vague, general, and of a more fundamental psychological nature. One tends to see far more use of "abilities" and "aptitudes" of the individual divorced from specific job content. After 50 years of disappointing validity coefficients, this approach certainly cannot be termed startlingly successful. And it may explain the increasing shift toward simulated prediction testing where job-related situations are used (for example, the Assessment Center).

Let me hasten to state that I am not implying an absolute dichotomy between past and future performance appraisal, and I am certainly not trying to complicate an already complex problem. I am not necessarily implying even that the two measures sets should be physically and temporally separated. I am suggesting that personnel action might be easier if they were kept separate--even if only in the same form.

What is really of importance, of course, is the use to which the data will be put. To reward for past performance is one thing; to fire, retain, or promote in the future is something else. Of particular importance here is t ncentive pay system of PL 95-454 ("Merit Pay and Cash Awards") ich is additional pay for past performance.

Money: Who Should Get It?

PL 95-454 partially replaces the current lock-step pay schedules with a more flexible merit pay and cash award system. To quote Chapter 54: "(B) use performance appraisals as the basis for determining merit pay adjustments." The Act does not specify the exact system to be used; the system is to be established by the Office of Personnel Management. But certain dimensions are specified:

> (i) any improvement in efficiency, productivity, and quality of work or service, including any significant reduction in paperwork;
> (ii) cost efficient;
> (iii) timeliness of performance; and
> (iv) other indications of the effectiveness, productivity, and quality of performance of the employees for whom the employee is responsible.

It is not clear to me whether or not this list is intended to be exhaustive. But the steps that have to be taken seem clear:

1. From the preceding guidelines specifically define the measurement dimensions and, possibly, add dimensions.

2. Define how these dimensions are to be measured (objectively, subjectively, without or with employee participation, etc.) and define what kinds of numbers will result.

3. Develop some sort of algorithms that will relate performance appraisal numbers to merit increases and cash awards.

The purpose of this Law is to provide flexibility and not anarchy. As a supervisor in this system who must do actual performance appraisals, I want some guidelines and techniques that are not only fair but that appear to be fair or, at the least, procedures I can believe are fair when I face an employee for the performance evaluation.

What is All This Going to Cost?

The Merit Pay system of PL 95-454 does not mean that more money will be available within the system. So, the Act itself does not necessarily imply increased costs for supporting Federal employees. But there are other "costs" to be considered. Three come to mind:

The Cost of Doing It

Any performance appraisal system takes supervisory and managerial time. As an example, using the present system, one supervisor used 110 hours (almost 3 weeks) of FY78 in performance appraisal for a 30-person group. This is about 6% of the total yearly time avaialble to the supervisor which in itself is neither a good or bad number. The question is: How much time and resources will the new system(s) cost?

The Cost of Appeals

One purpose of the new Act was to simplify adverse actions and the appeals procedure; in short, for example, make it easier to fire people. Whether it will or not remains to be seen. It is not, however, a license for arbitrary and unilateral supervisory behavior. For every adverse action, justification will be needed. How much time and supervisory effort this will take is a serious question.

The Psychological Cost

The general and positive purpose of PL 95-454 is to provide ". . . a competitive, honest, and productive Federal work force . . . and to improve the quality of public service." At the moment, however, the Act represents change and uncertainty to the present work force. This, in turn, may lead to hostility and resistance as specific procedures are introduced. At best, there will be confusion and apprehension which may lead to decreased productivity in the immediate future. An important contribution for psychologists to make (in addition to specific appraisal system methods) is techniques for introducing these systems with the least turbulence and the greatest possible acceptance. This, in fact, may be more important than the performance appraisal tools themselves.

Title VI: Research and Demonstration

One of the most interesting features of PL 95-454 is the provision (Title VI) for selected research programs and demonstration projects. The need for "improved methods and technologies in Federal personnel management" is specifically recognized as is the need to ". . . conduct and evaluate demonstration projects." For each demonstration project a detailed program plan must be approved by the Office of Personnel Management and submitted to the Congress. This type of demonstration is one for which psychologists are particularly (if not uniquely) trained. The psychological community in general and industrial/organizational psychology in particular should attempt to participate in--if not actually control and execute--these demonstration projects if there is anything useful to come from them. These demonstration projects in themselves may be more important than any of the specific tools we could design from them. They may be the largest controlled demonstrations of performance appraisal ever conducted.

PL 95-454: A Paradigm

The Civil Service Reform Act of 1978 provides an interesting and important real world human assessment problem. It impacts on thousands of individuals and hundreds of organizations. The new system that will be created provides a fundamental paradigm in which all the problems of human assessment and performance appraisal may be found. Like all human systems, it will in the future have some successes and some failures. In the process it may be an extremely useful case history against which new and old methods of human assessment might be evaluated. Since the Act itself encourages improved techniques and the research associated with their development, the evolution of the system may provide an unusual research opportunity--if researchers have an opportunity to develop and test innovations within the new system.

The pressure of time, however, is acute. Many features of the Act must be in place and operating in a very short period of time. To meet many of the dates (e.g., for the Senior Executive Service) much of the system will have to be designed on the basis of expert opinion. It is to be hoped that included in the heart of the initial systems will be incorporated sufficient measurement capability to assess the validity, reliability, and fairness of the system during the early days of operations. Perhaps a very high priority item at this time is to ensure that such a measurement system is included in the system from the beginning of its operation.

Dr. Muckler: Every time I come and visit the Air Force I'm either asked, "Do I speak for the Navy?" or I feel like I'm held responsible for something that the Navy did. Anticipating that, I've already got my disclaimers ready. It happened again. I do not speak for the Navy. I only speak for myself.

And can I join Mike here? I like to change my mind constantly. I think it was Oscar Wilde who said that consistency is the mark of a second class mind.

In my paper I took much interest in and focused on Public Law 95-454 and it's here. If you do not have it, I really would recommend to you that you get it. I have read it from beginning to end and I tried to apply for hazardous duty. Like all laws it refers to a lot of other laws and I think it may be a year's career trying to figure out if it is really basically changing several other laws.

I'm interested of course from two standpoints. One is, as a supervisor in this system I've got to face the consequences of what's coming from this. But I think moreso it is the biggest performance appraisal system forthcoming that I have ever seen. Starting on the first of July and by the first of October 1980, theoretically, whole new sets of performance appraisal systems will be in place covering 2,600,000 people. So we've got a very interesting paradigm here and it seems to me that I see going on here many of the problems that I've seen so many times before.

I'd like to explain to you just a little bit what happened, because there's a point to be made. Public Law 95-454 establishes the Senior Executive Service. And that's all the current GS-16's and above, of which there are about 3,000 in the Federal Government, and it has to be in place the 1st of July. One of the things that has to be done is that each agency has to recommend to the Office of Personnel Management a performance appraisal system for the Senior Executive Service. Now there has been some ambiguity in what an agency is. Our agency theoretically is the Department of Defense. Each service is charged to recommend a performance appraisal system to the DoD which will become, maybe, the performance appraisal system for the Department of Defense, subject to the approval of OPM, and possibly subject to the approval of Congress.

NASA has already submitted a plan to OPM. I just got a copy of it late Friday and didn't realize what I had till Saturday. We had to run off the copies this morning here and I'm afraid that some of the pages may be cut off. At the present time OPM considers this to be the bell weather system, and it may be because this is the first one they got in, and because they may be so happy to see one. But I think that what you see here is the kind of system we're going to be dealing with and the problem as I see it is that frequently the basic outline and the basic constraints of the system are decided before we get there.

I've frequently been called in when I was in industry. They'll say "We want you to approve our scales or our rating forms," and I look at their system and I say "I don't want to do that. I don't approve of your system because it's based on some false assumptions." I think that what we've got here is a system which may not be based on false assumptions but may be based on many questionable assumptions. The people who are going to pay the benefit of this is line management because we're going to have to take the brunt of it. There will be turbulence. I think what we're in for is a very substantial 2 to 3 years of much turbulence and much confusion. And I think, as you will see, the kind of open end system we've got here is sort of ripe for confusion.

I'm going to try to be as friendly about this thing as I possibly can. The key pages are 3, 4, 5, and 6. A basic procedure is described here which I think is sort of interesting and some of it which may sound pretty good and some of it which is really going to create some problems. First of all, the basic dimensions are defined and I think that's really interesting. There are three fundamental factors, if you will: management, the program, and the individual. So this is a three-factor system. And on page 5, the critical elements of each one of those factors are defined. So one of the problems that I raised in my paper, which I frequently have, is the deficiency of the performance dimension set. This system says on page 2 that this list of critical elements is a list of the total performance expectations. So page 5 defines a priority that these are the dimensions that we're going to rate SES people on. One of the ones in the law that they had to get in someplace was this "significant reduction in paperwork" and much is made of that in the law. They have that under "management performance" under "improvements in work or service." In my discussions with some of the people unfortunately (and I hope I'm going to the wrong folks here) but they don't understand . . . . I said, "If you're going to talk about reductions what are you going to do for the baseline measurement?" But that concept was new to them, I mean literally.

Those are some of the problems we're facing here. When I look at page 5, I say, "Is that complete? Is that indeed necessary and sufficient to describe the behavior?" These are high level supervisory people and I would ask you at your leisure, "Is this a satisfactory set of dimensions?" And again, this is paraded as being necessary and sufficient. This is complete. I'm not sure if you win a Nobel prize whether that counts in this system. It's certainly not an easy one, but I think they probably didn't worry about that.

Lt Col Ratliff: I'm sure you have raised a pretty good point. When we talked to the NASA folks we raised the issue, "What about the technical side of performance?" And they said, "Well these are top managers; they know everybody and they'll give credit due." Although that was basically their response, that doesn't provide documentation for the kind of competition that I understand that people of this level will be subjected to.

Dr. Muckler: I've already been talking to some of the people who are going to be in the SES, and they very much resent "individual initiatives." These are senior management people and they say, "I've got enough going to keep the shop going, and they want to do these kinds of things." If I might comment, the Senior Executive Service is going to be voluntary. The 3,000 or so people can go into it or not go into it as they wish. The first straw poll of those people showed that less than 50% may go into this.

Lt Col Ratliff: It is a fact though that everyone will be evaluated under whatever the SES evaluation program is for that agency, so they may not be gaining anything by staying out.

Dr. Mullins: There has also been some feedback from some fairly important people (I can't remember who it was now) saying that anybody who didn't opt for SES, if eligible, is not likely to have a fallback position in the long run.

Dr. Kavanagh: The law says . . . .

Dr. Mullins: I know what the law says.

Dr. Muckler: If they pull that kind of stuff, then I know they're going to see court action very very quickly. The law says these people revert to 15's, and a place will be made for them.

Lt Col Ratliff: That is if their performance is unsatisfactory only.

Dr. Muckler: No, no, I'm talking about the initial option that they have.

Lt Col Ratliff: I know, but they can remain as supergrade or as a member of the SES.

Dr. Muckler: So they are guaranteed. Pages 3 and 4 are the procedure, and this is MBO. The supervisor has to sit down with the person and in column 2 they have to come to two kinds of decisions within each one of these factors. One is this thing called a special objective. A special objective is going to be some new thing that you're going to do and you've got to have one for each factor, and then the continuing responsibilities are those things that you're doing already. So we've got two classes of events.

A really interesting column is column 3. I'm sorry that's cut off because that says something really interesting and it gets back to what you guys were talking about before. It says what the person has to do and what the rater (the supervisor) has to do as well. So you're really just establishing a contract here between the supervisor and the supervised. Now the question that comes to my mind right away is, "How well can people do this? How well can people sit down and develop such a contract?" My own experience with this has been rather poor. We have a narrative contract that they're going to develop, and it would be really rather interesting to see how well do people really do this. I think we're going to have a lot of embarrassing situations when they try to talk to each other.

Furthermore, column 4 establishes relative importance and that's a simple ranking. I'm not sure why they do that. I think it's a good thing to do but they don't use it again, but that's all right. Now, in 6 months and 9 months--and that timing is curious to me--column 5 is actuated and that's the actual achievement. How well did people really do? What results were achieved and how were they achieved? So you have to have a narrative description in there, and a four-point scale indicating the degree to which objectives were met, exceeded fully, partially, not at all. So here's some real precision measurement. There's a global rating with four points. And that'll be interesting to see how well it goes.

Once this has been done, then this is all combined on Table 6 where a summary rating is given in terms of the management factor, the program factor, and the individual factor, but it has to be one of the five categories at the bottom of the page. It has to be outstanding, highly successful, successful, minimally satisfactory or unsatisfactory. So we're making global judgments all the way along here.

What is interesting to me is that right after that you see the requirement for a general narrative summary. So apparently they're nervous here about that kind of global judgment and they want some words to back it up. My suspicion is that that's not enough lines to cover what's going on here. Then again it's got to be summed and an overall rating which has to be one of the five down below. You notice no guidelines about the algorithms to be used here. A realization that all of us have is that once you start accumulating this you get into some funny situations. Then at the executive review the person--and this is sort of cute--the person says I do or I do not wish to provide a written response and/or request review by a higher level official. So you will have an option, you will be able to write in response to this, or you can ask for higher review.

And the higher review processes are not too clear. Some of the talk around about this act is that it's going to give us a chance to fire people. And that's baloney. As far as I read it and as I see the appeal process, I'm just as scared to try to fire people as I was before.

We had an adverse action involving firing not long ago at NPRDC. The supervisor had to spend half-time for 18 months to get that action. And this guy was fruity. He was gone. He was sick and he should have been out of the system. That brings me up to a general comment I'd like to make about all the administrative and legal actions and really enforce what has been said before here. The employer is guilty until proven innocent. The supervisor is guilty until proven innocent. And some of these administrative hearings are so humiliating that it's beyond belief. I want to give you just one example. This guy didn't come in to work for 6 weeks and that was one of the items on the list of things. He was temporary and they didn't want to give him a permanent, which was very reasonable. The question was, "Did you tell him to come in to work?" And the supervisor says, "Yeah." And the next question was, "Did you give it to him in writing?" He said, "No." Then the administrator here says, "Well, that's it." It was really a humiliating experience for the supervisor and I think that particular supervisor may never bring an adverse action again. And this went on for hour upon hour, day after day. The self-esteem of the supervisor got lost in the system.

Let's suppose that the NASA system is adopted. It may well be. This may become the model system, and as I understand it the different agencies will be allowed some flexibility, but again that's not for sure. But let's suppose this is it and we get forced into using this. Before we've even got to play the game, many decisions have been made about the basic formats, the basic decisions, the basic ways of counting, and before we've even come into the situation. I think it's plain that one can see that the opportunity for litigation is just unending. And I think that we're just in a spectacular shape to litigate on this particular problem.

There's an additional deficiency here that's just staggering and I want to turn to that. The whole goal of this system is pay for performance. We've gone through all of this stuff and we've got a set of ratings on people, and they are one of these five categories of things. Then the next question becomes, "Who's going to get the money?" Well, first of all let me take up the problem about where's the money coming from? The law says very clearly we can't expect more money. We've got to do it within the existing budgets. There are only three places that this can come from. One is the cost of living increases and if they take that money then we're going to court again. That's a separate law and I don't see how they got away with that. The second is from our in-step increases, and that's a source of money. The third step has been the cash awards and the financial awards that are already being given. In our system nobody knows how much money that is. So we're not sure how much money is going to be available in the pot to give people.

But I want to give you an example of the kind of problems we're going to get into. A system like this has been tested in the treasury department this year for 700 senior level employees. Now they had $150,000 available to them so they decided that they would give 10% of those 700 people cash awards. I am unable to find out how they decided that it would be 10%, but we're going to have to have some sort of policy directorate. If you're evaluating as in our system 300 people--what percent are you going to have? I don't know how you solve that problem because then you can say well, whoever's good, and you get into a Catch 22 situation. But they decided 10% of the people would get cash awards. That's 70. And dividing that into $150,000, if you do it equally (and there's another question--what would the distribution be?)--that means a little over $2,000 per person. If you take off the income tax that's in that bracket they're down to about $1,100 and this turns out to be a very small percentage of their annual income.

So then you're going to get into the problem of when you give these kinds of increases will they mean anything to people? And we're going to have to get into the problem of perception of what pay means to people. A system has been created here that effectively is pay for performance. The rules for establishing merit pay have absolutely not been established and who's going to establish the rules has not been established. I think you can see why we are somewhat apprehensive about what we face.

Dr. Mullins: I might add one further comment. Fred has mentioned some of the difficulties with this law. If you'd like a few more interesting little problems with it before you get your copy of the law, sometime out of this structure I'll be glad to talk to you about it. It's pretty interesting.

Dr. Ree: It's very difficult to disagree with Fred about anything that he has said. So I perhaps will continue picking on the law. That seems to be a reasonable approach at this time.

What I really want to talk about is what I generally consider orthodoxy. That is, I remember somewhere reading that the Mandarin Chinese were the first to develop the civil service system, and they did that for one reason or another for their own orthodoxy at that time. Prior to the development of the system, conditions were such that they felt it was a good reason to implement this system.

In the 16th Century when the Ottoman Turks drove the Christians from the Holy Land, they established a civil service system that was based upon another orthodoxy. That orthodoxy was not the type of orthodoxy that we have today. It was a system in which you were not employed on merit, but by how much you could enrich the next person up the line. They developed this system of baksheesh where you paid money to the next person up. So I for example would collect fees or extort fees from the people who came to me for a service and I would be expected to pay a portion of that to my supervisor.

Well, surely that is not the orthodoxy today. I don't notice my supervisor living high on the hog. But orthodoxies change.

The original civil service act was based on the orthodoxy that there was a need for government services and that civil servants would be susceptible to political pressure. Now that was some hundred years ago, a little more than a hundred years ago, and that orthodoxy probably was founded then.

The prevailing orthooxy today is not that we're susceptible to political pressure, rather that we're lazy, inefficient, slothful, and we can't be fired. I don't know an example of any of those to be true en masse in Civil Service. I can certainly point to one or two people whom I personally consider are lazy. On the other hand perhaps they can point the finger back at me. This is the prevailing orthodoxy.

And this particular orthodoxy is based on another orthodoxy that I like to call merit. And, as Fred points out, what we need to do with the prevailing orthodoxy in 1979 is that we should reward merit (whatever that is), we should penalize the opposite of merit (and I'm not sure what that is yet--that may be mediocrity or that may be misfeasance) and, if necessary, we should be able to fire those people in civil service who are not meritorious.

In 1971 before I came into the civil service system I observed that if you did not like the president, he was limited by law to 10 years or two terms and I could outlive him. I've already outlived five or six presidents by this time. If I did not like my congressman or my senator, surely he was bound to change. They had limited terms of office. But at the time I was a government contractor and I had a monitor for whom I worked, who was not to me a meritorious individual. I found that I couldn't count on outliving him, and I couldn't count on his term of tenure expiring, and that perhaps what I should do was look for his chair. Maybe that's why I'm here.

Well, we have to consider the orthodoxy that talks about merit. So we have to ask, "What is merit?" and we have to ask a corollary question, "Merit to whom?" There are a great many people that may assess merit as we look at a system.

As I was growing up in the post-war years, having been born during the second world war, my generation in particular was told that one should be meritorious, one should study, one should apply, one should work, and one would then get ahead. That was the orthodoxy of the late 1940's, 1950's, and almost into the middle of the 1960's. By the time I entered the work world in the late 1960's I found that that orthodoxy like a great many others of them wasn't so much being challenged, but it wasn't true.

There are a great many things that enter into what permits an individual to achieve and what permits them to get ahead. I notably think of a friend of mine who went to a big time Ivy League graduate school in biophysics or something along that line, something esoteric. He struck me as an incredibly meritorious' individual. He was interested in plant physiology. He seemcd like a brilliant person to me, a brilliant man. Four years later when he got his PhD, he confided to me that he had studied the wrong molecule, and there was no practical application for it, and the best he could do was hope for a post-doctoral fellowship. Now, he was meritorious and it didn't work for him in the way that I view merit.

Those four ways are to the employee, to the employee's peers, to the supervisor of that employee, and to something we like to call management. When we think about merit as perceived by the individual, surely we've all looked at theories of satisfaction, we've all looked at theories for needs, etc., and we recognize that an individual may perceive merit in things that others do not. Certainly employees view themselves many, many times differently than do their supervisors. Perhaps that's why we allow for this interchange. That's why we want participatory evaluation. The peers of an employee may view merit still differ- ently. Merit to the peers of an employee may be a consequence of verbal footwork, may be a consequence of being in the right position at the right time, may be a great number of things which may or may not be related to performance on the job. I would surely like to know if we could assess that. Supervisors--it seems to me that in the final analysis, this is where the rubber meets the road to use Dr. Brokaw's metaphor of earlier. Supervisors in fact are the first source of power within an organization. They're not the only source but they're the first source. If you cannot convince a supervisor that you're doing a good job, surely he is not going to go to his supervisor or management (if that's what you want to call it) and he's not going to fight for you. So surely perception of merit by a supervisor is extremely important. My last one is management and I think to myself I don't know exactly what management is. I don't know what decision I've reached on this, but I see that at least we can look at merit a number of wa.'s.

I'm not sure that we've defined it adequately and of course the law does not define it accurately, but the prevailing orthodoxy is that we will reward merit, whatever that is, and we will punish the opposite.

In thinking of the opposite I'm not so sure that it's "dismerit." That's not the proper term. I thought about the concepts of the opposite of merit, malfeasance, nonfeasance, misfeasance, etc. Well it seems to me an individual may not be doing his or her job in certain areas and doing a terrific job in other areas. It may be that an individual is technically extremely competent but as a, for example, supervisor, they have no interest and they're incompetent, if you will. They're not doing the job, very simply. So merit can go down a number of different roads and I find myself wondering which of them is important, let alone which one I can follow.

We can look at the concept of productivity. Productivity is also a problem. It seems to me that there is an emphasis on trying to manage R&D on a productivity basis. I find it very difficult to countenance. I don't see it as exactly the same sort of thing.

Well, productivity, especially when we talk about certain parts of Public Law 95-454, is leaned on very heavily, and what we mean by it, again, is amorphous. We can look at the quality side; we can look at the quantity side. The law is ambiguous. To me the most problematic point of it, of course, is the merit pay question. It's not a technical issue. I don't think they're going to have any difficulty, to be honest with you, in removing the cost of living allowance part of it. I think that the law states that government employees' salary increases will be commensurate with the cost of living. Perhaps somebody can tell me more specifically how the law's written, but it seems to me that inflation last year was upwards of 9% and I know there is a pay cap of 5 1/2% next year. Surely that is inconsistent with the spirit of increasing the federal employees' salaries to keep up with the cost of living. So I don't doubt that they will change that too.

The thing that is to me the most problematic then is this 13, 14, 15 merit pay problem. And the most problematic part of all is the political part. Agency, as Fred tells us, is ill defined. Actually it's not defined at all. What an agency is is open to question. If agency means DoD, maybe I'm better off because I know we've got some people who will be in there fighting for the Air Force. If agency means Air Force, that means that various organizations under Air Force get to divide the money. That means that how well I fare depends on the adequacy and the forcefulness with which the Air Force Systems Command is willing to fight for the money. If agency means Air Force Systems Command, which is the big umbrella under which this organization functions, it means how forcefully is the Director of Laboratories going to fight to get money to AFHRL. I can't help but think about the concept of putting myself in the role of the Director of Laboratories. If I were, I can't help but ask the question, "Am I going to give more money to the missile development people or to AFHRL?" What do I have to show that looks like the sidewinder? What do I have to show that looks like the sparrow? I have hardly anything like that. Well, the political consequences of this law could be devastating. Until agency is defined, it's going to be moot. We're not going to know what's going to happen, and it could have a terrible effect, which brings me to my last point.

My last point is, what effect will this law have, period? For example, in the 1 through 12 level, the law does not make as sweeping changes. I don't know of any particular changes in the 1 through 12 level that are going to raise the hackles on anybody's neck. Looking at the 13-14-15 level, without addressing the SES (I never look that far into fantasy), looking at the 13-14-15, I see that if it is carried out in a way that removes the purchasing power from 13-14-15's, then those people are going to use the information they have in the most rational way. They're going to say, "If I can convert, if I can turn this around, if I can go to this company or that company, if I can go to so and so, that's where I'm going."

I'm a firm believer in the fact that people use information in a rational way. I think that one can point to a great number of examples in the market place, among them an unpopular model automobile, and on the other hand a best-selling model, and any number of things. And I believe people use information in a very rational way.

The consequences of this law could be that it has a very deleterious effect upon the federal civil service, or alternately it may be gamed and it may be changed just the same way as many other civil service systems. It may permit people to function the way they do function now. Those to me are the two possible alternatives.

Dr. Muckler: I omitted a rather critical point. This concerns Chapter 47 of the law--Personnel Research Programs and Demonstration Projects. Apparently in the writing of this law it did occur to somebody that this could not be set in stone and handed down with the tablets, and provisions are made for R&D and for these major demonstration projects.

I know we in the Navy already have two in submittal and one of ours will involve two laboratories with over 5,000 employees. So far one thing that's been omitted from the paper I've seen is evaluation and measurement. If there's anything I think we do well, it's that. So one of the problems we have is whether or not we should get involved in that and set up a measurement team for it, because their naivete about measurement is incredible.

What we're facing right now is a decision about how to use our R&D resources which I'm sure you are facing as well. So we're thinking about getting involved in all these demonstration projects, first because it's well worth doing, and second because I think we can help them a great deal. But the other part of this is R&D and we have been told that we will be expected to use our 6.2 and our 6.3 money, some of it, for research programs in this area and we have also been told that they have to be relevant to whatever system is picked. Since we don't know what system is going to be picked, it's making it a little difficult for us to plan. It makes me wonder whether the research that we have been doing, like much of the research we've been talking about today, is the most relevant to what we are facing here. I wonder if we may not be put in a position where we have to put our 6.2 and 6.3 money in these areas into research which may have to show very high face validity with the system which has been accepted.

Lt Col Ratliff: You're saying that your lab has been told by manaoement they're going to have to put 6.2 and 6.3 money on civilian personnel research?

Dr. Muckler: That is correct.

Lt Col Ratliff: We haven't been told that, outside of the RPR (Request for Personnel Research) that we responded to as a normal research reouirement.

Dr. Muckler: We have not officially been told this, but we might as well have. So it leads us to a very difficult position. Should we continue research like we've been doing or should we turn to research that we can show is quite relevant to the new system?

Lt Col Ratliff: Let me digress from your remarks there a little. We have heard that the Air Force has interpreted what was said in the Congress and from the testimony of the personnel people that they (Congress and the personnel people) do not believe that the various personnel systems of agencies in the government are competent to develop appraisal systems or to manage personnel. That the Air Force has been led to believe this has been told to DoD, and the network of the appointees at that high level who are working the problem do believe this.

We were told by the aerospace agency people, for example, that they deliberately excluded the personnel people except their own personnel director. He is a member of these various high-level committees, but the rest of the members are "professional" managers, and this agency and all of its centers are staffed about the same. So their system is a management generated system. I was told by a high-level official there that the personnel people would be bookkeepers for their system and it would be management generated, management developed, and management run. He pointed out that the system I briefed was a personnel system, developed by personnel people. And I said, "Yes, that's true." I said that the management people I briefed had a hard time understanding some of the concepts. The thing was, wherever we have turned or whatever feedback we have had from those high management levels, except one person, has been that management will develop the new system, and it will be a management run, management based system.

That is the ethos that seems to be permeating the system. And I'm not sure of the role personnel people will take. Now, if DPK (USAF Civilian Personnel) has already dedicated themselves to the system we're developing and looking upon us as a research resource, 2 years ago before the law and before OPM, etc., . . . They were looking at their problem as a significant problem when things were primitive--and they may have a personnel developed and run system.

But in terms of what is going to happen overall--Army is running another management-generated system, that is a supervisor rating form that is very complex. I understand now they're really emphasizing supervisor-worker interchange and interaction. I have in my folder some copies of the form that they were using 4 or 5 months ago and they ran a field test without really collecting data on it. I'm not sure how that's going to turn out. Dell Toedt and I spent quite a bit of time sitting and talking with the people who are working that problem.

Dr. Muckler: This goes back to what Malcolm was saying about past and current orthodoxies. I sense from our management and by that I mean our politically appointed management, a spectacular hostility and I think if we try to present factual assistance, factual data, I'm not sure they'd listen, and I'm not sure there is any set of words that they're going to hear that's going to make any difference.

Lt Col Ratliff: The thing that I've heard at the Pentagon talking to people at various levels, is that there is a circuit of the top people who continually work these problems over the telephone, in meetings, and who get very incensed about details and frustrations. Their views are the only perceptions apparently that are prevailing. They're not looking for professional inputs. They're looking for political and management solutions to what they perceive to be a management problem and are not really concerned with the technology. They're looking at the politics; i.e., who is going to control the review committee that the law says will be established, how much military membership can be tolerated, etc., etc.? And some of the interchanges among different elements are becoming quite vehement I understand.

Apart from that I don't think the Air Force for example has worked that problem (the political problem), and nobody is worked up about it. There probably is some of this lying under the surface but it hasn't come out as a problem to be recognized by management yet. It's something that is being dealt with by people on the Air Staff for the SES group saying, "We will have the workers and supervisors together in training. We will have this form and we will have that set of conditions, and we will give you the solutions." I don't know really where that part of it stands at that high level. I don't know what the Army is doing about this specifically.

Dr. Muckler: There is always a search for simplistics. "If I give someone a merit increase, it will increase that person's productivity."

Lt Col Ratliff: In fact, one high level manager told me that they were going to pay the good people more and it would really work well. And I asked, "How are you going to find the good people?" And she said, "I see what you mean."

Dr. Ree: Well, this is the prevailing orthodoxy, that you pay peoole more and they'll work better. I don't know if there's any evidence at all that supports that. They just work more expensively. We've tried things like suggestion systems or outstanding performance awards. Some of the studies indicate that those don't really motivate behavior. They seem to create hard feelings among the people who didn't get them for a couple of weeks and then everything returns to the way it was. That is if you believe the orthodoxy of wanting to reward merit with that.

There may be other reasons that you do that. There may be any number of reasons why we do things, and a rewarding of merit may not be among them. And perhaps justly so. There are occasions when we don't want to do that.

Dr. Muckler: In that environment I'd offer them the Mark Twain principle. There's a simple solution for every complex problem, and it's always wrong.

CHAPTER 7

MEASUREMENT AND LATENT-TRAIT THEORY

By

Malcolm James Ree
Personnel Research Division
Air Force Human Resources Laboratory
Brooks Air Force Base, Texas

Often the characteristic to be assessed is an inferred ability or trait. Although it is impossible to directly observe such concepts as mathematics ability, or knowledge of general science, or tool knowledge, it has become common practice to add up the number of test items answered correctly and infer an examinee's ability or, more precisely, to infer some location for the examinee on the continuum of ability. The use of measurement techniques based on the number-right (or corrected-for-guessing) score was facilitated by the development of true-score theory and its associated test and item statistics.

True-Score Theory

Traditional true-score theory calls for the analysis of items on the basis of two well-known parameters: difficulty (*p*) and item-test correlation coefficients (*r*). Item difficulty is usually defined as a function of the number of examinees responding correctly to an item. Item-test correlations indicate the discriminating power of the item as a function of the higher probability of correct responses by examinees achieving higher scores. In the dichotomous case, items are scored as correct or incorrect. Usually a 1 is given for a correct response and 0 for an incorrect response, and the subject's score is the sum of the item responses. In the polychotomous case, weights are usually assigned to each answer category such as in the familiar equal interval scaling where, for example, like = 1, indifferent = 0, and dislike = -1. Instead of a *p* value, the item mean is computed and item-test correlation is computed in the usual manner, and subject scores are computed as a sum of the item responses.

In both the dichotomous and polychotomous case, the observed item analytic indices have the common fault of not being indenendent of the group of examinees sampled. For example, *p* value is dependent on the average ability of the group. If the distribution of the scores for the group has a large skew, the *p* value will be shifted, or if the test scores are collected on a group with a leptokurtic distribution of observed scores, the true-score item-analytic indices collected on one group may not be applicable to another similar group that is not as leptokurtic. These problems have been well known for quite some time.

As early as the second world war, Tucker (1946) proposed an advanced theory of measurement to overcome problems with true-score theory but was unable to pursue it due to the lack of automated computing facilities. Lord (1952) further refined derivatives of these ideas and proposed the basics of a modern theory of measurement. Again, investigation and implementation of the theory would await high speed computational facilities. Birnbaum (1958) developed an advanced model based on the general shape of a curve which related correct item response probability to an examinee's ability. Implementation of this theory awaited the availability of fast and efficient digital computers because the model is mathematically complex and computationally laborious.

This modern theory of measurement provided for solutions to problems encountered in classical true-score theory. It allowed for measurement of ability independent of the composition of the test. This permits direct comparisons of scores for different examinees administered tests composed of different items within a single ability domain by relating ability to the probability of answering an item correctly. Other advantages of the theory will be discussed later in this paper.

Latent-Trait

For convenience, the measured trait will be called ability and denoted by theta ( ) although other mental traits could be measured accurately by using this latent-trait model. The applicability of this model to performance ratings, or to interest or vocational measurement, is fairly direct. Dichotomous items or ratings can be analyzed or scored with the appropriate model as can polychotomous or scaled ratings, thus bringing the advantages of latent-trait theory to these areas.

Latent-trait theory is based on the probability of an examinee's answering a test item correctly as a function of that examinee's ability. The relationship between the examinee's response and the unobservable trait, say arithmetic computation, is described by a mathematical function.

Three salient features of the general theory of latent-trait measurement must be described. These are: dimensionality of the latent space, local independence, and item characteristic curves.

The number of traits which underlie examinee test performance is described by the dimensionality of the latent space. It is customary to assume that the latent space is unidimensional, which is equivalent to the assumption that the test items measure but one factor. The same assumption is made when classical reliability is estimated through item homogeneity methods such as KR-20. Although many tests are believed to violate this assumption, recent data indicate that its effect is not sufficiently detrimental as to render the model unworkable.

Local independence means that an examinee's performance on one item does not influence the responses to other items. Specifically, item responses are assumed to be a function of only ability and no other extraneous factors such as race. This is in effect a restatement of the assumption of unidimensionality of the item pool. If the item response is only a function of a single trait, then the relationship between item responses must be independent of extraneous factors. On the other hand, if the item responses are related by some additional factor such as racial experience common to only one group, then local independence is not obtained and the item pool cannot be unidimensional. Something other than the latent-trait is being measured.

Item Characteristic Curves (ICC) show the regression of item response probability on the latent-trait. One way to distinguish among the extant latent-trait models is to note the mathematical form of the ICC.

Figure 1 shows three ICCs. The ICC for a linear model is shown as 1-a; Figure 1-b shows the three parameter logistic ICC; and 1-c shows the item-option response curves ("Nominal Response Curves") for a single item. Empirical evidence indicates that the linear latent-trait model does not adequately describe examinees' responding to aptitude test type items (Urry, 1977a). This model has the simplest mathematical form.

$$P(\theta) = b_g + a_g\theta \qquad (1)$$

where P (θ) is the probability of marking a correct answer and a_g is a function of the slope of the ICC line and b_g is a function of the item difficulty.

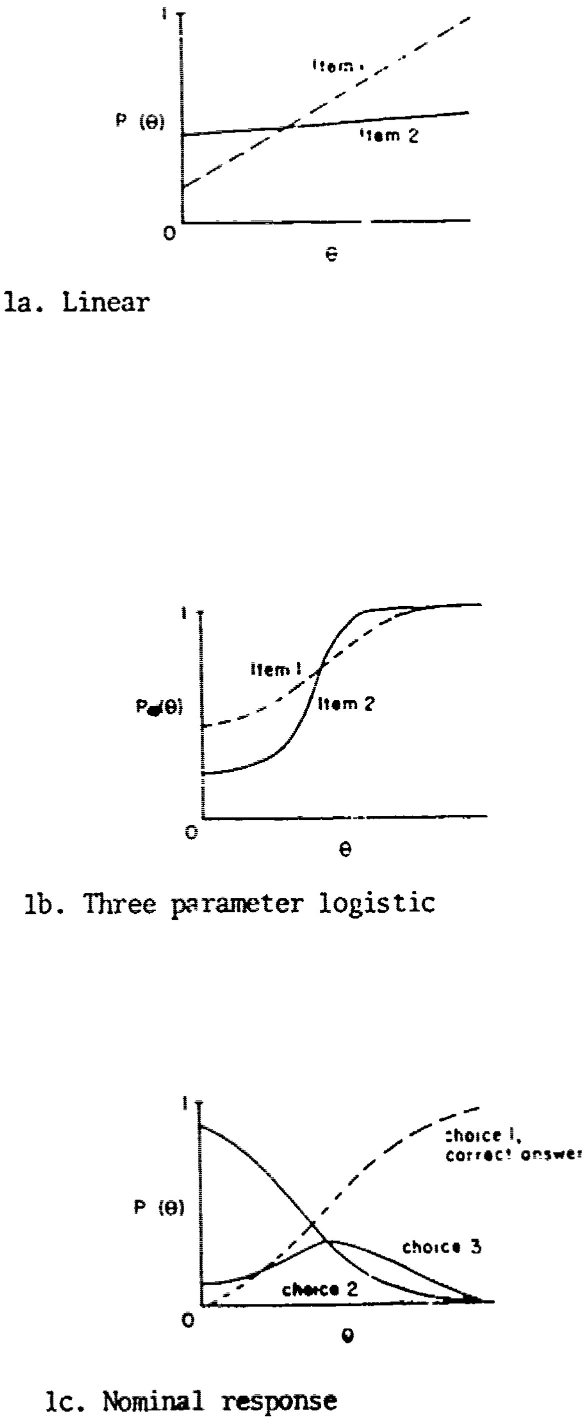
The Birnbaum (Lord & Novick, 1968) three parameter logistic model is the most frequently used for relating item responses to subjects' ability. The three parameters, a, b, and c, are item discrimination, item difficulty (or location), and probability of chance success (or lower asymptote), respectively.

The curve described by these parameters takes the shape of an ogive (cumulative frequency) or an "s" with the upper asymptote approaching a probability of 1.0 and usually a lower asymptote of a probably greater than 0. The ogive describes the probability of obtaining a correct answer to an item as a monotonic increasing function of ability.

The item discrimination parameter (a) is a function of the slope of the ICC and generally ranges from .5 to about 2.5. The value of a equal to about 1.0 is typical of many test items while a values below .5 are insufficiently discriminating for most testing purposes, and a values above 2.0 are infrequently found.

The item difficulty parameter (b) describes the point of inflection of the ICC and is usually scaled between -2.° and +2.9 with a mean of 0.0 and unit variance although the metric is arbitrary. The b parameter describes the ability level at which one-half the examinees answer the item correctly and is scaled in units of θ.

The item guessing parameter (c) is the lower asymptote of the ICC and is generally conceived of as the probability of selecting the correct item-option by chance alone. Most test items have c parameters greater than .0 and less than or equal to .30.



1a. Linear

1b. Three parameter logistic

1c. Nominal response

Figure 1. Examples of Latent-Trait Curves

The curve describing the Birnbaum three parameter logistic model is given by:

$$P(\theta)_j = \underline{c}_i + (1 - \underline{c}_i)\ (1 + \underline{e}(-1.7\underline{a}_i(\theta - \underline{b}_i)))^{-1} \qquad (2)$$

where $P(\theta)_j$ is the probability of "subject" j answering test item i correctly and a_i, $\underline{b}_i$, and $\underline{c}_i$ are item parameters for item 1 (Lord & Novick, 1968).

Figure 2 shows three Birnbaum type ICCs. The horizontal axis is scaled in units of ability, and the vertical axis is the probability of answering the item correctly. The solid curved line shows an ICC for an item of average difficulty with acceptable discrimination and the lower asymptote appropriate for a five-option multiple choice item. The dashed line shows an item of identical difficulty, $\underline{c}$ value of .28, but with a lower $\underline{a}$ value. Note how the slope of this curve is less steep. The third curve, dot-dash line, shows an item with a $\underline{c}$ value of .30, an $\underline{a}$ parameter of 1.0, and the $\underline{b}$ parameter equal to 1.0. As the $\underline{b}$ parameter changes, the location of the inflection point of the curve is displaced along the horizontal axis.

The Nominal Response model (see Figure 1-c) is used to describe response probability to each item-option of any one item. Each item-option is described by a separate item-option curve. The general equation is:

$$P_{gk}(\theta) = \frac{e^{b^*_{gk} + a^*_{gk}\theta}}{\sum_{h=1}^{m} e^{b^*_{gh} + a^*_{gh}\theta}} \quad (g = 1, 2, \ldots, n;\ k = 1, 2, \ldots, m). \qquad (3)$$

The sum of the probabilities of selecting each of the item-options is equal to one for each level of . The Nominal Response model seems most appropriate for scaled rather than scored items and should be very effective in interest inventories or employee rating systems. For a further explanation, see Bock (1972) and Samejima (1972).

Because the Birnbaum three parameter logistic model is so frequently used, it will be discussed for the remainder of this paper. It should be noted that the characteristics of latent-trait theory apply to the other models as well.

Parameter Estimation

In most cases the test constructor is faced with the task of estimating three parameters for the $\underline{n}$ items and one ability parameter (θ) for every examinee (N) so that N + 3n parameters must be estimated for each group of test items. For a group of 2,000 examinees taking 80 items, 2,240 (2,000 + (3 x 80)) parameters must be estimated simultaneously. In an iterative procedure this estimation must be

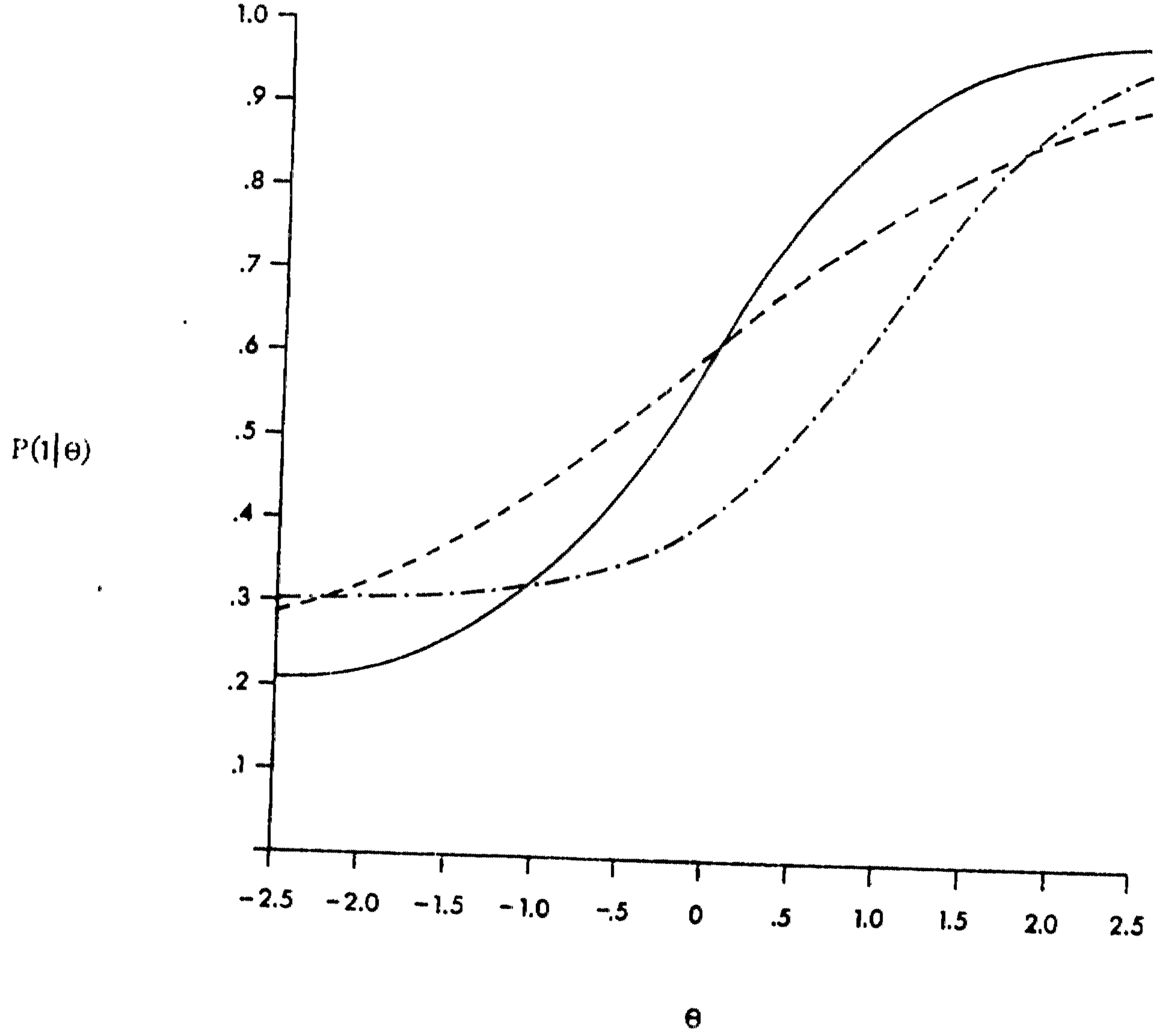


Figure 2. Item Characteristic Curves

repeated several times which leads to long computer runs with more precise estimates. Simulation studies indicate that latent-trait item parameters may be estimated very well. The b parameter is estimated with the most accuracy, a the next most, and c is estimated with the least accuracy (Ree, 1978a). It should be noted that the b parameter has the greatest influence on computing latent-trait estimates of ability while the c parameter has the least (Ree, 1979). Accurate estimation of item parameters may be done with no more difficulty than estimation of classical item parameters.

Estimation of Ability

There are several methods of estimating the subject's ability. The three most frequently used are: raw score, Maximum Likelihood, and Owen's Bayesian estimation. The last two are latent-trait based procedures.

Raw scores have the problems of (a) variability due to difference among sets of items, (b) variability due to choice of subject group to be tested, and (c) relatively poor regression on ability. These problems are avoided through the use of latent-trait estimation of ability.

Maximum Likelihood Estimation (MLE) of θ is computed using the likelihood function defined as:

$$L(\theta) = {}_{\pi}(P(\theta)^{u}\, Q(\theta)^{1-u}) \qquad (4)$$

where $Q(\theta) = 1 - P(\theta)$, u is 1 if the item was answered correctly, 0 if answered otherwise, and the product is across all items answered. The maximum of the distribution of likelihoods is found by the method derived by Jensema (1974). The use of this procedure is advantageous because it allows the estimation of θ regardless of the sequence of item administration. Other methods, such as Owen's Bayesian estimation of θ, are sequence dependent (see Sympson, 1976).

MLE is not sequence dependent but has the problems of possible failure to converge and convergence on an infinite estimate. Both of these problems can be rectified by arbitrarily placing a limit on the number of iterations and by placing an upper and lower limit on θ. Thus, θ may be estimated from item responses when the ICC parameters have been estimated. MLE also has a highly linear regression on ability over the entire range of ability (Maurelli, 1978) which facilitates accurate estimation at any level of θ.

Bayesian estimation procedures have been extensively studied (Jensema, 1972; Maurelli, 1978; McBride & Weiss, 1976; Owen, 1969, 1975; Urry, 1971). They avoid the problems associated with Maximum Likelihood estimation but tend to have a non-linear bias in estimating θ. Low θ's are frequently significantly overestimated and, as with any Bayesian procedure, there is the phenomenon of regression toward the mean which distorts the estimates.

Item and Test Information

Precision of measurement in true-score theory is based on the concept of reliability. A single value is used to describe a function of error variance (due to restrictive assumptions), even though it is generally acknowledged that error varies by ability level and item quality. Latent-trait theory offers an analogue which makes fewer restrictive assumptions and avoids the problems associated with reliability estimation. The precision of measurement, or reliability analogue index, is called Information, and it avoids the restrictive assumptions of true-score reliability. It may be thought of as the (reliability) precision of measurement at a specific level of θ.

Item information is defined as:

$$I_g(\theta) = (\frac{\partial}{\partial \theta} P_g(\theta))^2 \Big/ P_g(\theta)\ (1 - P_g(\theta)) \qquad (5)$$

where $P_g(\theta)$ is estimated from equation (1), and the numerator is the squared first derivative (i.e., the squared slope) of $P_g(\theta)$ at a fixed value of θ. Test information is the sum of the item information curves making up a test and is defined as:

$$I(\theta) = \sum_{i=1}^{n} I_g(\theta) \qquad (6)$$

where $I_g(\theta)$ is defined in equation (5) and n is the number of items.

It is useful to calculate item and test information curves in order to determine the precision of measurement of a test or an item. The height of the item or test information curve at any specified value of θ may be thought of as being an ICC analogue to classical reliability at that value of θ. The higher the information curve the greater the information value and the higher the reliability of the item or test and, hence, the greater the precison of measurement at that value of θ.

Figure 3 shows the information cur e for an item with an a equal to 1.2, b equal to .0, and c equal) .20. Note that the curve is unimodal and skewed, as is typical of most test items.

Figure 4 shows information curves for five items of identical a and c values but of differing b values. The dashed line which is above the individual item information curves is the sum of the item information curves and is the test information curve. Prediction of measurement at any θ value may be determined for a test by reading test information curves.

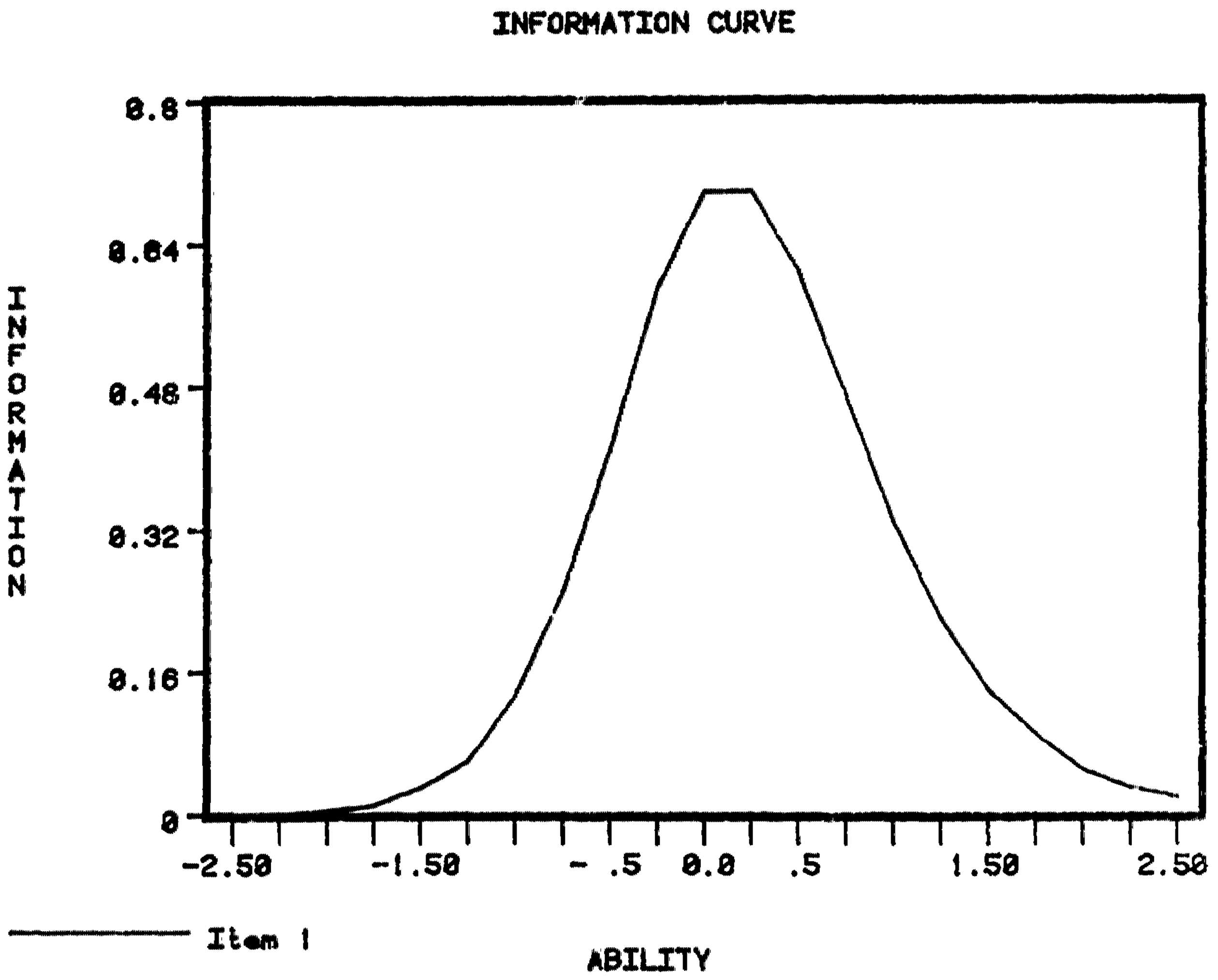


Figure 3. Item Information Curve for a Typical Item

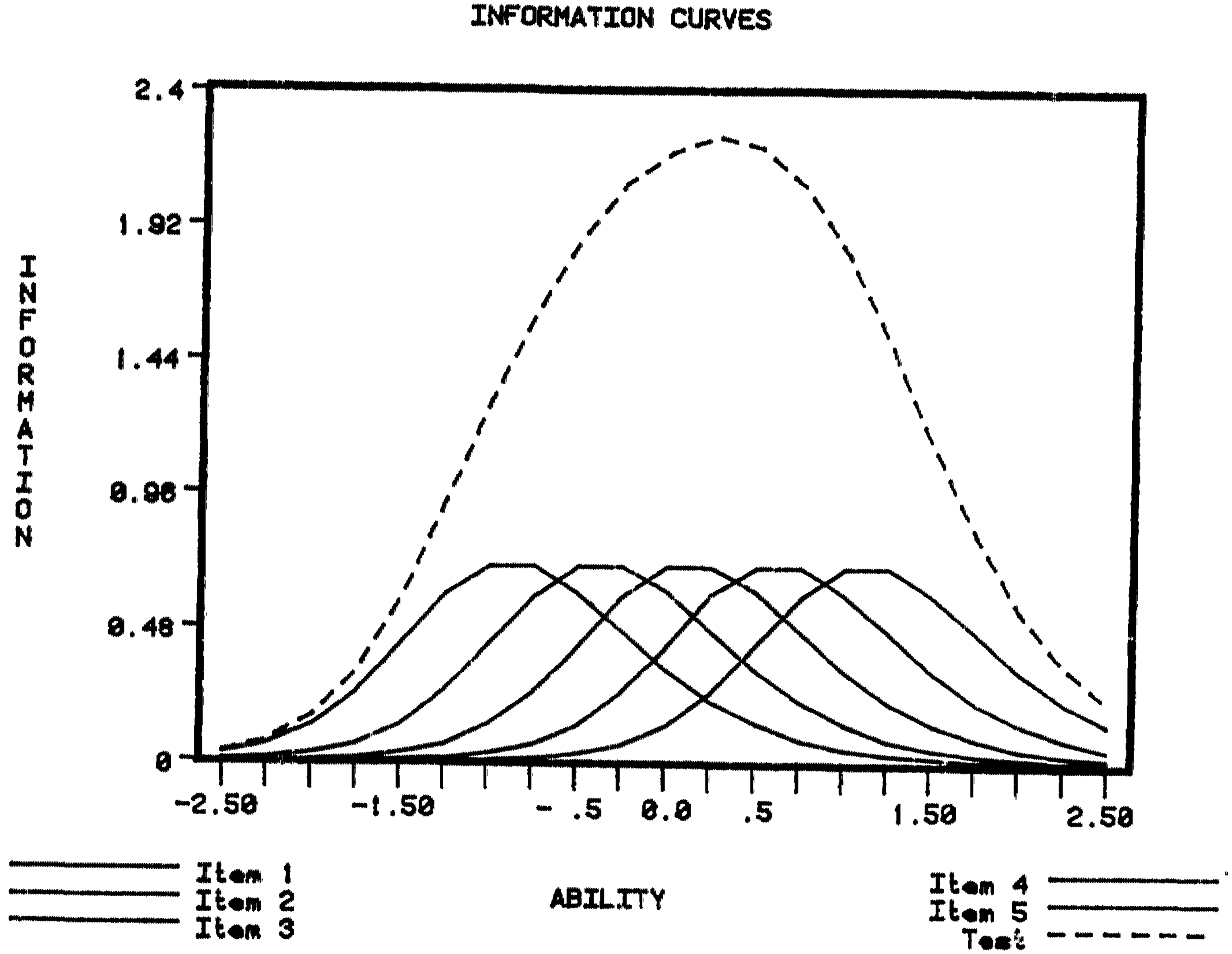


Figure 4. Test and Item Information Curves

Figure 5 shows the test information curve for a subtest from a typical multiple aptitude battery which is used for selection, classification, and counseling. Although decisions are based upon scores in all ranges, it may be seen that information is high only in the range from -.8 to +1.3. This has the practical consequence of lowering the validity of decisions made from outside the range of high information. It should be noted that true-score reliability estimates do not permit this type of knowledge.

Applications of Latent-Trait Theory

The results of simulation studies indicate that increases in test reliability and validity may be achieved by scoring paper-and-pencil tests by Maximum Likelihood estimation of θ. Ree (1978a) demonstrates an increase in test reliability for 80 simulated items from .939 for scores derived by number right scoring to .948 for scores derived by use of Maximum Likelihood estimation of ability for the same 80 items.

Item analysis and selection for test batteries can also be improved via use of the ICC item parameters. Jensen and Valentine (1976) report the construction of a short test for the prescreening of applicants for military enlistment which was developed using latent-trait theory. The test is used to select applicants who have a sufficiently high probability of achieving a score above the cutting point on the military enlistment qualification battery. These applicants are then provided with transportation to an Armed Forces Examining and Entrance Station (AFEES) and also provided with meals and lodging while there. This prescreening test was built by computing ICC parameters and selecting items which had _b_ parameters clustered around the desired cutting score. The use of the prescreening test can effectively reduce costs for recruiting and enlistment processing.

Item and test information curves can be used to make tests maximally discriminating at various cutting points. For example, if a minimum cutting score at the 20 percentile is required, then the items may be selected to produce a distribution of information peaked at this level of ability. Similarly a multipeaked or flat distribution of information may be constructed as required.

The use of latent-trait theory facilitates the automation of test construction. Ree (1978b), in developing an automated test item banking/test construction system, used the three parameter model to estimate test mean score, standard deviation, reliability, and percentile core equivalents. A study of the techniques indicates that the estimated test statistics were very close to the observed test statistics on a different group of subjects. On-line item storage and interactive test construction using latent-trait theory afford the test constructor accurate knowledge of test statistics before the test is administered in an operational setting.

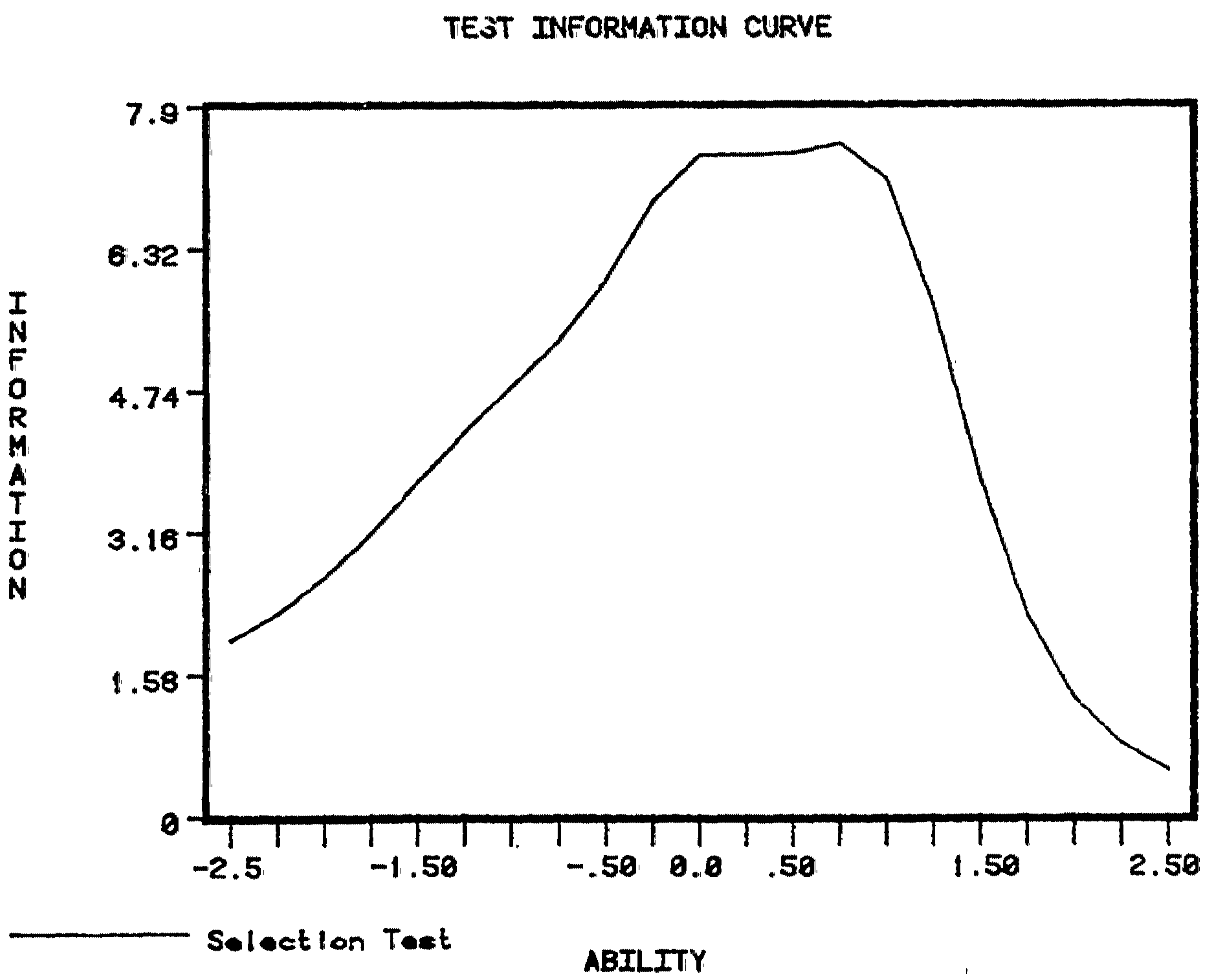


Figure 5. Test Information Curve for a Typical Selection and Classification Test

Lord (1977) demonstrated that ICCs may be useful for the selection of items for a conventional test, redesigning an existing test, equating test scores, and for adaptive testing.

Adaptive Testing

The greatest achievement facilitated by latent-trait theory is adaptive testing. This rubric describes a series of strategies for adapting the difficulty of test items to the examinee's ability level. Each of the strategies has the same objective, the improvement of the psychometric properties of test scores by adapting item difficulty during testing.

During adaptive testing, the item to be administered is selected based upon the response to the previous item (or responses to all the previous items), the response scored, ability estimated, and the next item to be administered selected. The number of items may vary from examinee to examinee, and there is a very low likelihood that two subjects would ever be administered exactly the same items. This would have the advantage of greatly reducing problems associated with test security and compromise, as well as the advantage of administering a test uniquely adapted to the examinee.

The feasibility of applying adaptive testing techniques to test batteries and to test items requiring both alphanumeric and graphic display has been demonstrated by Ree (1977). Three aptitude areas: Word Knowledge, Arithmetic Reasoning, and Space Perception, were administered to about 200 subjects at the San Antonio, Texas, AFEES, and demonstrated that adaptive testing using off-the-shelf technology was possible. To date this is the only instance of the presentation of graphic or pictorial items via a computer.

Brown and Weiss (1977) conducted a simulation study with previously collected item responses from 365 Naval fire control technicians. A total of 232 achievement type items were administered on a paper-and-pencil test which was divided into 12 subtests, each covering a different content area. These responses were used to simulate subject responses in an adaptive testing procedure. The 232-item test was also scored by conventional means and the conventional scores correlated with the adaptive testing scores. The observed correlations were above .80 for 11 of the 12 subtests; the 12th was .74. In all cases, these high correlations were achieved by adaptive tests using about one-half of the total number of items in the conventional subtest. By selective administration of items, the adaptive tests achieved a precision of measurement as high as conventional tests which were twice as long.

Sympson and Ree (in press) have studied the validity of two types of adaptive testing procedures on a sample of military technical training students. Two adaptive tests and one traditional linear test of Arithmetic Reasoning (AR) were administered to 490 subjects at Chanute AFB, Illinois. The adaptive tests were Bayesian and Maximum Likelihood, and for the sake of reduced computer effects the linear test was also administered on the terminal. Final course grades served as the validity criterion.

It was found that under specific conditions the validity of the adaptive tests was higher than the validity of the conventional test. The Maximum Likelihood procedure was found to be slightly more valid than the Bayesian procedures. Also, the adaptive tests showed a higher level of information for most examinees than did the conventional test.

Urry (1977b) has investigated via simulations several important aspects of adaptive testing. Among the most important is his procedure for administering a "multidimensional" adaptive test. This is the analogue to administration of a paper-and-pencil multiple aptitude battery. He developed procedures that permit the user to specify the level of validity required and asymptotically approach it by administering the appropriate number of items from each of the subtests.

Adaptive testing also provides a level of test security which cannot be achieved by paper-and-pencil testing, and test security has become an increasingly important factor. Adaptive testing items are stored on computer disk and are unreadable to anyone without access to the files, which may be locked to all but a few "privileged" users. It is also possible to encrypt the stored items and make them unreadable to all but those with access to the file and the proper decoding key. It would be extremely costly to achieve this level of security with paper-and-pencil tests.

Extensions of Latent-Trait Theory

The latent-trait model may be extended to other areas of human assessment. The three parameter logistic model has been proposed for use with a vocational interest inventory. Further, Samejima (1972) has proposed a graded response (scaled items rather than scored items) which appears to have a great potential for scoring interest inventories or questionnaires.

The application of either the three parameter logistic model or the nominal response model to employee, task, or merit rating systems is a logical extension of latent-trait theory. For example, if the items were of the type

Check one of the following:

Completes work

a. before it is due.
b. when it is due.
c. after it is due.

then the nominal response model (see Figure 1c) might be used to determine the rating for the employee. Such a rating system would then have all the advantages now found in latent-trait based aptitude tests.

Another extension of the theory is creating person characteristic curves (PCC) rather than item characteristic curves. These curves, which would be shaped like a normal or logistic ogive, would relate the probability of a specific person's responding to items (or ratings) in a particular manner as a function of the characteristics of the item. For example, Figure 6 shows a PCC for a specific individual. It is determined by estimating the individual's θ and then computing the probability of obtaining a correct answer to questions already asked. The theoretical PCC is the solid line, and it may be compared to an observed PCC (dashed line) when sufficient items have been presented at each level. Cumulative differences between the two curves would serve to indicate non-standard administration conditions such as coaching or random responding. This same approach might be applied to personnel ratings by plotting PCCs for one supervisor and comparing it with a PCC for another supervisor. If the two (or several) supervisors are measuring the same trait but with different units of measurement and different origin, then one rating can be shown to be a linear transformation of the other, and a latent-trait model would apply. This would allow one person's rating to be directly equated to another's regardless of the number of rating items or the relative scale any individual supervisor uses. The PCC is a recent development which, although holding great promise, requires considerable research.

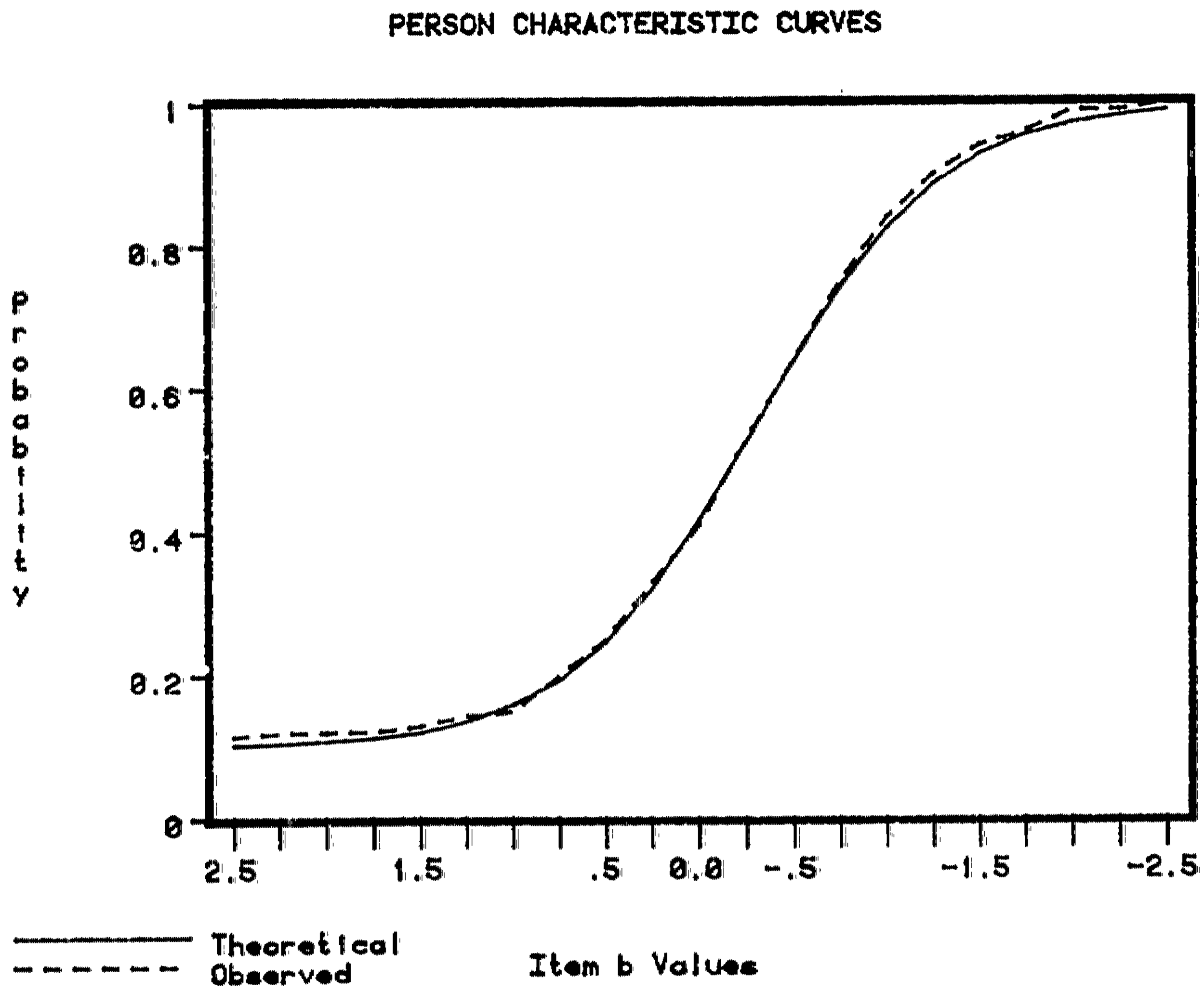


Figure 6. Person Characteristic Curves

REFERENCES

Birnbaum, A. On the estimation of mental ability. Series Report No. 5. Project 7755-23, USAF School of Aviation Medicine, Randolph AFB, TX, 1958.

Bock, R.D. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 1972, 37, 29-51.

Brown, J.M., & Weiss, D. An adaptive testing strategy for achievement test batteries. Research Report 77-6. Minneapolis, MN: University of Minnesota, 1977.

Jensema, C.J. An application of latent-trait mental test theory to the Washington Pre-College Testing Program. Unpublished doctoral dissertation. University of Washington, 1972.

Jensema, C.J. An application of latent-trait mental test theory. British Journal of Mathematical and Statistical Psychology, 1974, 27, 29-48.

Jensen, H.E., & Valentine, L.D., Jr. Development of the enlistment screening test-EST Forms 5 and 6. AFHRL-TR-76-42, AD-A033 303. Lackland AFB, TX: Personnel Research Division, Air Force Human Resouces Laboratory, May 1976.

Lord, F.M. A theory of test scores. Psychometric Monograph, 1952, No. 7.

Lord, F.M. Practical application of item characteristic curve theory. Journal of Educational Measurement, 1977, 14, 117-138.

Lord, F.M., & Novick, M. Statistical theories of mental test scores. Reading, MA: Addison-Wesley, 1968.

Maurelli, V. A comparison of Bayesian and Maximum Likelihood scoring in a simulated stradaptive test. Unpublished thesis, St. Mary's University, San Antonio, TX, 1978.

McBride, J., & Weiss, D. Some properties of a Bayesian adaptive ability testing strategy. Research Report 76-1, Minneapolis MN: University of Minnesota, 1976.

Owen, R. A Bayesian approach to tailored testing. Princeton, NJ: Educational Testing Service, Research Bulletin RB 69-92, 1969.

Owen, R. A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. Journal of American Statistical Association, 1975, 70, 351-356.

Ree, M.J. Implementation of a model adaptive testing system at an Armed Forces Entrance and Examination Station. Proceedings of the 1977 Computerized Adaptive Testing Conference, Minneapolis, MN, July 1977.

Ree, M.J. Estimating item characteristic curves. AFHRL-TR-78-68, AD-A064 739. Brooks AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory, November 1978a. Also published in Applied Psychological Measurement, 1979, 3, 371-385.

Ree, M.J. Automated test item banking. AFHRL-TR-78-13, AD-A054 626. Brooks AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory, May 1978. (b)

Ree, M. The effects of errors in estimation of item characteristic curve parameters. A paper presented at the Meeting of the Military Testing Association, San Diego, CA, October 1979.

Samejima, F. A general model for free-response data. Psychometric Monograph, 1972, No. 18.

Sympson, J. Estimation of latent-trait status in adaptive testing procedures. Proceedings of the 18th Annual Convention of the Military Testing Association, Gulf Shores, AL, 1976.

Sympson, J., & Ree, M. The validity of a Bayesian and a stratified Maximum Likelihood adaptive test. In press.

Tucker, L. Maximum validity of a test with equivalent items. Psychometrika, 1946, 11, 1-11.

Urry, V.W. Individualized testing by Bayesian estimation. Research Bulletin 0171-177. Seattle: Bureau of Testing, University of Washington, 1971.

Urry V.W. Tailored testing: A spectacular success for latent-trait theory. Springfield, VA: National Technical Information Service, 1977a.

Urry, V.W. Tailored testing: A successful application of latent-trait theory. Journal of Educational Measurement, 1977b, 14, 181-196.

Dr. Ree: Let me start my presentation which is somewhat different from the rest of them. Let me talk to you about what I am about.

The term latent-trait theory is somewhat unfortunate. The concept of a latent-trait conjures up any number of mystical notions about deep hidden recesses of the mind and things like that. It is rather the unfortunate naming of a theory of measurement that was proposed in the 40's by Ledyard Tucker, prior to the development of high speed computers.

I'm going to try to develop the theory just a little bit for you inasfar as it's necessary to understand my proposed nexus of this theory to ratings research. My paper begins by looking at what we call classical or traditional item analysis. One of the things we notice in item analysis is that the indices by which we judge the items are not independent of the other items that fall into the measurement device. Notice I haven't said "test" here. I'm including rating scales and things like that, and the indices are not independent of the people upon whom the indices have made the measurement, whether it be test items or rating scales. For the sake of clarity or for the sake of my convenience, not necessarily yours, I'm going to speak in terms of test items because that's the way I'm used to talking about this.

In the 50's, or perhaps after the second world war, we developed something known as "true score theory," and a number of people who worked in the Army Air Corps psychology program later went on to write books on this particular theory, notably people like Gulliksen and Cronbach and Thorndike, any number of people. They were all part of the Air Force psychology program. It was a roster of who's who in psychometrics and, in some degree, personnel selection, etc. Well, after the war, Tucker noticed that there was this problem of the variability of these indices. That is, they were not invariant, depending upon who you gave them to. Now he proposed a method that he called the "constant process" and it was based upon a shape of a curve relating the probability of getting the item correct to the score that the individual got on the particular test. Or, to put it into rating terms, the probability of selecting a rating category to the overall rating, if I may make that sort of an analogous jump. That didn't go very far because it was computationally laborious in 1946, 33 years ago, to compute the things we had to do.

In 1952, Frederick Lord who is now the chief psychometrician at one of the large commercial testing companies developed a theory that he called latent-trait theory. He proposed in a monograph, I think it's Monograph #7 in the Psychometric series, an extension of Tucker's theory. He doesn't give Tucker very much credit but if you read the two of them you'll see the similarity.

In 1958, about 6 years later, the Air Force engaged Alan Birnbaum, who has since died, to develop a modern theory of mental measurement. That modern theory was supposed to remove this subject variability of the items and the item effect on the subjects. That is, how can we ask two people different questions? That's not fair treatment. But the fact of the matter is that it very well may be fair treatment. It may very well be necessary to ask different people different questions in order to make proper measurements for those two individuals. Well, Birnbaum came up with a three-parameter model that bears his name, also called the logistic model, and I will discuss those three parameters but I think that it's more important to give you an overview of what a latent-trait model looks like.

We have endeavored somewhat in this community to call this "item response theory." It's a nicer name. In fact I had a work unit here where I was looking at the accuracy of our ability to compute these curves and I called it something about latent-trait theory and they sent it back from the headquarters shop saying we don't do anything at all clinical here. And I agreed with them and I changed the name to "item characteristic curve" and the same thing went through.

The curves that are shown here are in fact item characteristic curves and they deal with this concept of measuring some unknowable trait. That's why it's called latent. If I ask you what your temperature is, well you can measure that directly. You can take a thermometer of some sort with some accurac, and you can measure it and you can say to me, "Cookin' at 75 today; that's a little slow," or you can say 102 or 98.6 or whatever it is. If I ask you, for example, "How good are you at mathematics? How good are you at arithmetic reasoning? How good are you at word knowledge?" you can't tell me in the sense that you can give me a scale value. All I can do is give you a series of items, perhaps a very long series of items to make some estimate of your location on that ability. That ability, because we can't measure it directly, is therefore latent; it's unobservable, and that's why it's called latent-trait theory or item response theory today, if you like.

The particular model shown here shows different ways of conceptualizing what the relationship between marking an answer and overall ability would be. For example, we can take a very simplistic model and we can say that it is represented by a straight line and you can see item 1 and item 2. And over along the vertical axis we have the probability of that response, let's say the probability of getting the item correct, as a function of the individual's ability, which for convenience we just call "theta." That's ability. We'll talk about how we might measure that in a moment because these don't spring full grown from the brow of Zeus as did Athena. These are measured in some way or estimated in some way.

Well, here's item 1 and it's characterized by a straight line. This is the kind of thing we implicitly do when we calculate item-test point-biserials, or biserials. We simply say "Oh yes, there's a low group down here and there's a high group up here," and we bang a regression line through them and the hell with the rest of it. What we should be doing perhaps, is we should be looking for the shape of that regression and when we find the shape of that regression in an item that is scored correct or incorrect, typically it takes on the shape of an ogive, it takes on a logistic shape. That is, it takes on the shape of a normal or a logistic cumulative ogive. It starts somewhere, it has some lower asymptote, it has an inflection point right at the center here where it changes acceleration and then it goes up to another inflection point.

Dr. Mullins' growth charts, in fact, could be modeled after this, although let me point out one difference here. These are assumed to be monotonic increasing functions of ability. That is, the better a person is, the more likely they are to get the question right. Put into ratings, the better their ability in something, the more likely they are to be rated highly.

Dr. Cascio: You used the word "logistic." What do you mean by that?

Dr. Ree: Logistic. It means that it is a log transformation. One can demonstrate without getting into the mathematics of it, that the difference between the logistic ogive and the normal ogive is never greater than some scaling factor. And since we can know that scaling factor, we can make fewer assumptions in the mathematics of it.

So the logistic curve just has an exponential function, if you like. What you find is--if, instead of assuming a linear realationship in doing your item-total correlations, if you scatter plot those instead of assuming a linear relationship--I've been playing around with that a little bit and I found--that a quadratic function fits it better, a cubic even better than that. Don't forget that a cubic function--I like to think about it this way--the number of inflection points tends to be one less than the number of degrees in the polynomial; therefore, a 1 has got no inflection points and a cubic of course has 2, and this is a cubic, more or less. A quadratic looks slightly different but we have to make different asumptions because a quadratic says, "Yes, we can turn down here toward the end of the curve." We don't want to do that.

We don't want our ratings and our measurement to be such that somebody who is better has a lower probability of getting that choice on the item. If we can scale it correctly, what we do instead of having the mean of the lower group and the mean of the upper group as we do in traditional item analysis, how we do it is we divide this into intervals and we get the probability of passing at that interval and regress that against the interval value. So if we're down here at -2, we find that the probability of passing is .20. We're down here guessing somewhere.

Dr. Cascio: That's a standard scoring unit?

Dr. Ree: They are whatever units we care to make them. But yes, they could be standard score units. They look like standard score units, but they could be anything you like. Scaling is arbitrary. We talk about Evaluation by Objective (EBO) and having 10 points for this and 20 points and we multiply and get 200 points. That's a fine scale and it's not intrinsically any better than multiplying it by 2,000. It's an arbitrary scale and it has its values for us.

The one that I think has the greatest application to rating work is what is called the nominal response model, which is perhaps the most general case where each item option or each response category has a known characteristic function and each of these curves that are shown here is the function for that category, given that we fit this particular curve and find that for various parameters that are necessary we can compute a likelihood function as to where that individual is, given that we get enough items. And enough items is an empirical point.

An overview of the model says that we want to relate probability of being chosen in that category, of getting the answer right, or whatever it is to some location on this parameter which we call ability. Given that, we're offered several advantages. I will talk about it in more detail, but some of those advantages are that (1) we can tell when ratings are not working at all, or I suspect we will be able to tell when ratings are not working at all, and (2) if we have strictly a leniency error (that is, an adding of a constant), if we have one supervisor who rates everybody with a mean of 5 and another one who rates everybody with a mean of 4, if in fact one is simply adding a constant to the other, we should be able to determine that. There is a way of anchoring all those ratings into a common metric, and that to me is one of the possible advantages we would find in here.

I want to talk about the parameters that are usually used to characterize this particular system and in doing so I have three curves up here. These are basically for test items rather than for scaled items. However, conceptually, this may be one of the scale points in a rating item.

The three things that we use to characterize items are: (1) the lower asymptote of the curve. In a scored item as opposed to the scaled item, that can be shown to be the probability of getting the item correct by guessing. That is, with no knowledge, what does a person do? Do they guess? That is the lower asymptote, so that we have what we call the *c* parameter here which is the guessing probability on the item. We find through experimental work that that is not always equal to the reciprocal of the number of item options. The fact of the matter is that there is a large body of research that shows that people guess in a way that tends to be like random but in fact is not, because we use partial information in a number of things we've all seen before. (2) we don't want to give an item to somebody who may be down here if the item is working up here; we don't gain anything from that. That's sort of like my asking you to add 2 and 2, and just consistently asking you to add 2 and 2, 3 and 3, 4 and 4. I cannot tell anything about your ability in calculus, for example, or your ability in arithmetic reasoning from that. So there is some reason, for example, to move items to where a person is rather than the other way around. It is an item location parameter. You may think of that as a rough analog, if you like, of the item difficulty parameter in classical analysis. (3) there is that something that may be considered in a nontechnical sense the instantaneous slope of the line at the inflection point. It is an item discrimination parameter. Remember, I said there was an inflection point on the curve. The inflection point is in fact the point of maximum discrimination on that particular curve. Discrimination in the sense of distinctions; that is, where this particular item is working the best. Well, that inflection point tells us about where on this ability scale the item is working maximally. It is something that says, this is how steeply the curve rises and this is how quickly it makes distinctions among the individuals.

Guttman has proposed, for example, a latent-trait model. Guttman items have all these characteristics, and one thing you notice in Guttman items is that the slope is entirely vertical. You know it or you don't know it. Well, we all know from our elementary psychophysics at least that--things could be a step function. We should see something or we shouldn't see it. But in fact, they never are. They take on this kind of logistic form. In fact, this was reported in the biology literature 50 and 60 years ago. They called it lethal dose 50. That's what the *b* parameter was here. This is the dosage at which 50% of the organisms died. So, you gave a drug and 50% of them died and that's how you characterized that particular drug.

Those three parameters--the guessing parameter, the item location parameter, and the item discrimination parameter--those particular parameters are invariant to a linear transformation of the scale of ability. That is, it makes no difference to the shape of this curve if I label this as -3 and this as -2 and so on and so forth. I can grab hold of this metric if you like and wiggle it back and forth and it doesn't change the shape of this curve. That has some nice advantages, very very nice advantages.

Let me talk to you about what we've done with this and try to relate it to ratings. We have looked at our ability, for example, to estimate these particular parameters. We've looked at this through simulation studies and we can do that fairly well, and we have looked to see which of these particular parameters is easiest to estimate and which is the most difficult to estimate and we know which of those we can estimate well. Fortuitously, the one that is the easiest to estimate, that we can estimate the best, is the one that's the most influential. Once we've estimated these parameters, the next step in determining what this item is doing for us is to go down and find some way of estimating ability.

When I gave this paper to some of the members of the staff here to read they said to me, "Where do you get this ability? Is this something mystic that you sat down one day and plugged some numbers in the computer and came running out with two tons of paper and said here it is?" Well, the answer is simply, "No." The answer is that we make an estimate of ability by using certain known functions that tell us about the way this curve looks.

Typically we do it in one of two ways. There is something known as Bayesian estimation of ability or Bayesian modal estimation of ability. Who was it that was mentioning Vern Urry to me? Was it you Wally? Vern is a very big proponent of Bayesian modal. There is another method we use, a maximum likelihood method. And what we do quite literally is pull ourselves up by our own bootstraps. But the fact of the matter is, strangely enough, that it works because we do it in an iterative fashion.

If, for example, I were to start out down here with z scores as you suggest, I can take the a, b, and c parameters of those z scores, go back to my original vector of item responses, and through a couple of iterations arrive at a stable estimation of theta. The fact of the matter is that I have demonstrated through simulation studies that we can do this very quickly and with a great deal of success. So we can estimate theta by making crude estimates of it. Those of you who are familiar with maximum likelihood estimation, for example if you've seen the Newton-Rapheson procedures, you can put nonsensical weights in there and eventually, although it takes longer, it will iterate down to the true value. So we can do that sort of thing here.

One of the other advantages of this theory is something I should like to introduce here, an item information curve. Let us stipulate that the item information curve that I'm showing you here is simply from one of those options on the rating scales. I'm not nearly as facile with terms like BARS and so forth, that's not been my main consideration for the last 10 years. Let's just look at a BAR and ask how measurement is done. For example, let us note that a bar has a rating. Let's suppose this is simply one item option. What the information function tells us is how that item is making discriminations at any point along the ability spectrum. Suppose you've got to rate an individual on some characteristic, and suppose you've used your BARS system or whatever it is, and that you're able to scale the items as I propose. Well, I would suggest to you that we can find exactly where that particular item is working I would be able to tell you not only where it's working but relative to all the items that we can put on the same scale (a very big caveat), I can tell you how well it is making distinctions and among which individuals.

Now I think that I have skipped over a very important point. And that important point is the concept of unidimensionality. Very frequently we make an assumption of unidimensionality without ever noticing it. True score theory, which by the way is a latent-trait theory, tends to be a unidimensionality theory. Anytime you calculate a coefficient alpha or KR20 or something like that, you're implicitly making the assumption of unidimensionality. For if you're not, then you're wasting your time doing that sort of thing. This is based on unidimensionality. That, to me, is one of the things that one has to assess as to whether this will be applicable, or whether we need a multi-dimensional model. If we can get away with a unidimensional model, it's here. If we have to have a multidimensional model, we are faced with a large number of problems.

Dr. Borman: Malcolm, can you look at the individual categories separately and treat them unidimensionally?

Dr. Ree: That very well may be a possibility. Again, nobody has done any of this and I'm way, way out there in left field speculating and thinking about what the future may hold.

Dr. Borman: For instance, interpersonal skills versus administrative skills could be, it seems to me, treated separately as latent traits. You have two dimensions in that particular system, but . . .

Dr. Ree: It might actually be easier to treat them as one but we don't know. We have to look at it. This is all very speculative. I have not one shred of evidence that this is the way we should go on ratings but I think we ought to look at it, for sure.

Dr. Kavanagh: Two comments. One is that when I got to that point in your paper, I wrote in my notes "job performance is not unidimensional." I'm not sure I'm right, but I wonder if what Wally just said might also apply. The second thought I had is that Bob Guion gave a paper on this. Are you aware of the work that he has been doing on this?

Dr. Ree: No, not in the least.

Dr. Kavanagh: I don't know whether it's rating data or is it . . .

Dr. Bernardin: Both. He uses a three parameter model to look at test bias and a one parameter model to select behavioral statements.

Dr. Ree: If you would write down a reference for me, I would truly appreciate it.

Dr. Bernardin: It's Wright, 1977. I think it's a master's thesis.

Dr. Ree: Benjamin Wright is a psychiatrist. Actually, he's an interesting character.

Dr. Bernardin: No, I'm referring to another Wright. He did his work at Bowling Green, under Guion.

Dr. Kavanagh: I think it's all non-published at this point. I think it's a graduate thesis.

Dr. Ree: Well, I'm glad somebody is working on it. I'm only sorry we haven't heard about it yet. I don't claim proprietary rights to any of this. I'm simply saying that I think this is something we could look at. I hope to look at it within the next couple of years. The item or item option information curve, we can then use to assess the accuracy of our measurement by accumulating several of them. We can do this with abilities testing. We do this quite as a standard practice now with abilities testing here. We can take each of the particular items (you can see that they all have a maximum at a different point), and we sum them and we in fact can get the point where the test or where the rating scale is maximally informative to us. Which really ought to lead us to ask the question, if this item isn't working terribly well for this individual over here, why are we bothering with this item? There's no sense in it to me. Why are we doing that? Well, maybe we shouldn't be. Maybe we should be building our rating scales, our tests, whatever it is, in such a way that we are getting the maximum bang for the buck, where we are getting the maximum amount of information about the individual. Figure 5 shows an abilities test, and you'll notice that this is the actual calculated information function or information curve. And from about there down, about -.60 down, we're pretty shy on information. We're asked to make judgments all along here which we think we might be able to improve by amending the shape of this curve by redistribution of our items. Now what I propose is that this theory be extended into rating measurement. I'm glad to hear Guion has done it, but I'm sorry he hasn't let us all know about it.

Dr. Cascio: The first thing that popped into my head was that when you described that test information curve, it seems to be doing a better job of discriminating, if you will, at certain ranges than at others. It is exactly analogous to heteroscedasticity.

Dr. Ree: Well, I don't find that inconsistent at all. I like to think of it as a rough analog to reliability, the advantage being that reliability, which again when calculated by KR20 or anything like that, is a point estimate of something that is perhaps better not described as a point estimate. If you've got to make a decision about whether to take an individual into the Air Force for training as a weapons mechanic, and if we have reason to believe that that person should be at the 84th percentile on electronics, we really don't care what the reliability of the test for electronics is at the 20th percentile. That's totally alien to what we've got to do. We've got to make a distinction at a point that is well above that. Reliability theory always makes the assumption that we have this nice distribution of errors. Well, it's just not so. It doesn't appear to be so. Anytime we restrict an area we can demonstrate that it is not so. I'm rushing to meet our schedule if I can but let me talk about one other concept which is very new that we are trying to promote.

Dr. Borman: Let me ask one question before you do that. Is it true that to build these curves you need a number of raters rating a number of ratees in common? You need a complete crossed design in order to build these curves?

Dr. Ree: I don't know. I can't answer that with any assurance because I've not tried to do it in practicality.

Dr. Borman: It seems to me you would. I mean it could be done, somehow. It could certainly be done in certain situations, but in other situations it would be really hard to do.

Dr. Ree: It may turn out to be a large sample technique. I have no way of assessing that until we actually go out and try doing it. But that may be a failing of it, for all I know.

Dr. Borman: It wouldn't necessarily be a failing in a large number of situations.

Dr. Ree: Let me make another suggestion. The particular theory that we're looking at here proposes that people act in a certain way and one of the things that we can do is, given that we can make some estimate of an individual's ability, we might then be able to draw information curves across many bits of information on an individual. Remember, the information curve we have been discussing is one bit of information across people. This is across items or rating devices for one person. Now we can do a great number of things with this.

In terms of ability estimation, for example, suppose I get a very low probability of someone answering an easy question correctly, and yet they answer difficult questions correctly; i.e., the curve doesn't fit a monotonic increasing function of ability. Well, that very well may be an indication of coaching, test compromise, any number of things. We can make a maximum likelihood estimate of the probability of the person answering a question correctly. And inasfar as that deviates from the subject's observed pattern we may have evidence of nonstandard administration.

Another thing. Keeping the same idea in mind, let's suppose that we imbed a set of anchor rating scales for a group of raters, and those individuals through simulation techniques, through watching videotapes, or whatever it is, rate what they saw there using items of known characteristics. If we have people that are simply shifting the metric around, we can then take that individual's responses and through the invariance property of the three parameters that work on this, we can move eveybody on to the same metric. This, to me, was one of the things that I'd hoped to gain from the use of item response theory. This is all hypothetical and conjectural at the moment.

Dr. Cascio: I wonder how that would tie in with being able to get a handle on individual theories of conceptual likenesses.

Dr. Ree: I don't know.

Dr. Mullins: Wouldn't that likely reveal itself in different shapes of that curve rather than just moving the whole curve backward and forward?

Dr. Ree: That's an interesting point because now here we run into another problem that perhaps can be looked at. It can't be answered, but it can be described now better.

For example, if you have one rater who constantly gives a higher rating as a function of greater status of that trait, whether it's industriousness or whatever it is, and you have another rater who has a different shaped curve, it's obvious those people are not rating the same things, and it's obvious that you ought not to compare people who've been rated by those two individuals. It's obvious that it's unfair to compare people who have been rated by those two individuals when considering for promotion, if those ratings become part of the promotion system.

So I don't know. What you're pointing out may be a deviation from the model or it may be another advantage the model gives us, or if we develop a multidimensional model we may find that we may want to correlate on only one particular dimension or another.

Dr. Bernardin: You see any problems with estimating theta using ratings?

Dr. Ree: No, conceptually not. But conceptually bumble bees can't fly. So I don't know what to say to that. John, you may have a point that we can't estimate theta, given ratings, but we won't know until we try it, and I think we'll be ashamed if we didn't try it.

Dr. Mullins: I think you could definitely estimate theta using anything we want to. Whether or not that estimate is very good is another matter. I don't know how important that is.

Dr. Ree: Inasfar as these estimates are good, then the shape of the curve can be estimated with a great deal of confidence. However, inasfar as these vary, then of course the curve becomes fuzzy and we can't do it.

Dr. Kavanagh: This is in a particularly difficult problem area that we finally got around to addressing. I think in your paper it's designated as comparability across raters. There have been a number of schemes attempted in the literature to deal with this problem. It strikes me that part of the failing of those schemes is that you have to have a supervisor in a situation for such a long period of time that it's almost impossible to get enough good data. I wonder if this particular approach using a videotape type situation might be a solution. But that makes the additional assumption that observation and evaluation made in the taping situation is the same as that made on the job, and that may not be true. It might be a way of estimating errors. Carrying this through to its logical extension, that means that I would be taking some of my raters' data in the system and manipulating it when it comes into my computerized personnel system such that the ratings that they make will in some way be altered when it goes into the system. You see, that's the rub.

Dr. Ree: It's obviously not as open as simply permitting John to rate Joe and then Bill to rate Sam and then comparing Sam and Joe on ratings. I certainly like the idea of openness, but can we set a value to openness and a value to fairness and compare the two of them. It may be that openness is more important. Perhaps that is the hidden agenda in all of . . .

Dr. Borman: What you could do is just lay these ogives on the different raters and show them we're adjusting their ratings.

Dr. Ree: That's tough to explain. That's even worse to explain to the poor worker in the blast furnace who is sweating there, and you walk in and you say, "Well, we have this curve and this ogive . . .," and the first thing that worker's going to do is run off for the shop steward.

Dr. Kavanagh: The last point that you made is that, in order to get a good estimate of theta, different people should be responding to different test items. Therefore, to get a measure of true performance of different individuals doing the same job as described in the job description, I should really be assessing different pieces of information on each one. That appeals to me conceptually, and also in a sense that it creates a more open system--in that the feedback I give to individuals about what they're doing on their job is much better--and then it runs right into EEO guidelines.

Dr. Ree: The only thing that might be of interest here is that inasfar as we can compute person-characteristic curves, we might use it as an analytic tool, but inasfar as we can compute them, we might find that the person-characteristic curve for John when he's rating Blacks is different from the person-characteristic curve for Bill when he's rating Blacks, or for John when he's rating Whites, etc. We're going to look at this in terms of test bias that way. It seems to me that the seminal, the eschatological definition of bias in a test item or a rating scale is that some people with the same ability do not have the same probability of passing the item. If someone has a lower probability simply because you can identify their race, you've got a biased item. I think this could also be demonstrated for rating items.

My presentation is highly theoretical. I think there may be some good points to this. I think that we may find it's a large-sample technique and it may not be so good. But there are other things that we can look at, and I think it would be a shame if we passed it up.

SUPPLEMENTARY

INFORMATION

A102755

AIR FORCE HUMAN RESOURCES LABORATORY
Brooks Air Force Base, Texas 78235

ERRATA

Number	First Author	Title
AFHRL-TP-81-20	Mullins	AFHRL Conference on Human Appraisal: Proceedings

1. On page 181, the second line under the heading "Latent-Trait" reads "denoted by theta ( ) although ...." Pencil the Greek letter Θ inside the parentheses.

2. On page 184, the second line of the fifth paragraph reads, "item-options is equal to one for each level of ." Pencil the Greek letter Θ in the blank before the period.

LOU ELLIOTT
Chief, Technical Editing Office